Dense Non-Natural Sequence Peptide Microarrays

for

Epitope Mapping and Diagnostics

by

Joshua Amos Richer

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2014 by the
Graduate Supervisory Committee:

Stephen Albert Johnston, Chair
Phillip Stafford
Neal Woodbury
Antonia Papandreou-Suppappola

ARIZONA STATE UNIVERSITY

December 2014

ABSTRACT

The healthcare system in this country is currently unacceptable. New technologies may contribute to reducing cost and improving outcomes. Early diagnosis and treatment represents the least risky option for addressing this issue. Such a technology needs to be inexpensive, highly sensitive, highly specific, and amenable to adoption in a clinic. This thesis explores an immunodiagnostic technology based on highly scalable, non-natural sequence peptide microarrays designed to profile the humoral immune response and address the healthcare problem. The primary aim of this thesis is to explore the ability of these arrays to map continuous (linear) epitopes. I discovered that using a technique termed subsequence analysis where epitopes could be decisively mapped to an eliciting protein with high success rate. This led to the discovery of novel linear epitopes from *Plasmodium falciparum* (Malaria) and *Treponema palladium* (Syphilis), as well as validation of previously discovered epitopes in Dengue and monoclonal antibodies. Next, I developed and tested a classification scheme based on Support Vector Machines for development of a Dengue Fever diagnostic, achieving higher sensitivity and specificity than current FDA approved techniques. The software underlying this method is available for download under the BSD license. Following this, I developed a kinetic model for immunosignatures and tested it against existing data driven by previously unexplained phenomena. This model provides a framework and informs ways to optimize the platform for maximum stability and efficiency. I also explored the role of sequence composition in explaining an immunosignature binding profile, determining a strong role for charged residues that seems to have some predictive ability for disease. Finally, I developed a database, software and indexing strategy based on Apache Lucene for searching motif

patterns (regular expressions) in large biological databases. These projects as a whole have advanced knowledge of how to approach high throughput immunodiagnostics and provide an example of how technology can be fused with biology in order to affect scientific and health outcomes.

DEDICATION

I dedicate this thesis to my wife Sara for providing love, support, and all the most important things in life during this time

and

To Dr. Daniel A. Smith for teaching me to never let anyone hold you back from your dreams

ACKNOWLEDGMENTS

There are too many individuals who have influenced me in a positive way to thank here. Specific to the research presented in this thesis, Drs. Stephen Albert Johnston and Neal Woodbury have continually inspired me with their vision, drive, and unique way of approaching the scientific process and looking at the world. Their most valuable lessons were not specific facts or pieces of knowledge, but in showing by example the right way to ask questions, analyze a problem and find creative solutions to answer those questions and solve those problems. Dr. Bart Legutki also inspired me with his methodical work ethic and ability to pick up new concepts in a quick, nonchalant, deadpan manner that echoes his dry sense of humor. Dr. Phillip Stafford taught me many things about the practice of statistics, and how to balance the need for the ideal statistical study with the constraints of reality. Antonia Papandreou-Suppappola provided much needed mathematical expertise, and also an outside, unbiased voice on committee. Xiao Wang, Dr. Bart Legutki, Dr. Zbigniew Cichacz and the Peptide Array Core collected much of the raw high throughput data upon which this thesis relies. I also thank the Center for Innovations in Medicine, the Biological Design Graduate Program, as well as Achievement Awards for College Scientists (ARCS) for providing funding support during this time.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

The Need for Preventative Medical Technologies

The rise in healthcare expenditures has historically and consistently outpaced general inflation (Staff, 2013) despite record research expenditures aimed at better and more efficient treatments. This suggests a failure in existing paradigms in delivering care to an ever larger population. Instead of identifying disease early thereby making existing treatments more efficient, the focus is on creating new treatments for late-stage disease, resulting in a very poor return on investment for dollars spent on care, research and development. Healthcare spending in the United States is approximately 17.9% of GDP, the highest in the rich world (Martin, Lassman, Washington, Catlin, & Team, 2012). This systemic problem has stabilized in recent years at this level, but in the words of former President Bill Clinton "we are ahead by a country mile" compared to the rest of the rich world in terms of just how much we spend on healthcare. Despite this enormous burden, outcomes are only middling. This unacceptable situation has been met with only lukewarm political intervention and the situation shows no signs of improving in the future. Only new technology and more importantly, new ways of thinking about healthcare, can slow or reverse this trend. This chapter outlines the problems and trends facing modern healthcare and healthcare research, advocates for a new approach based on prevention and early detection, reviews the literature on antibody based diagnostic technologies that could meet this need, and outlines the specific contributions of this thesis aimed at addressing a small part of this larger challenge.

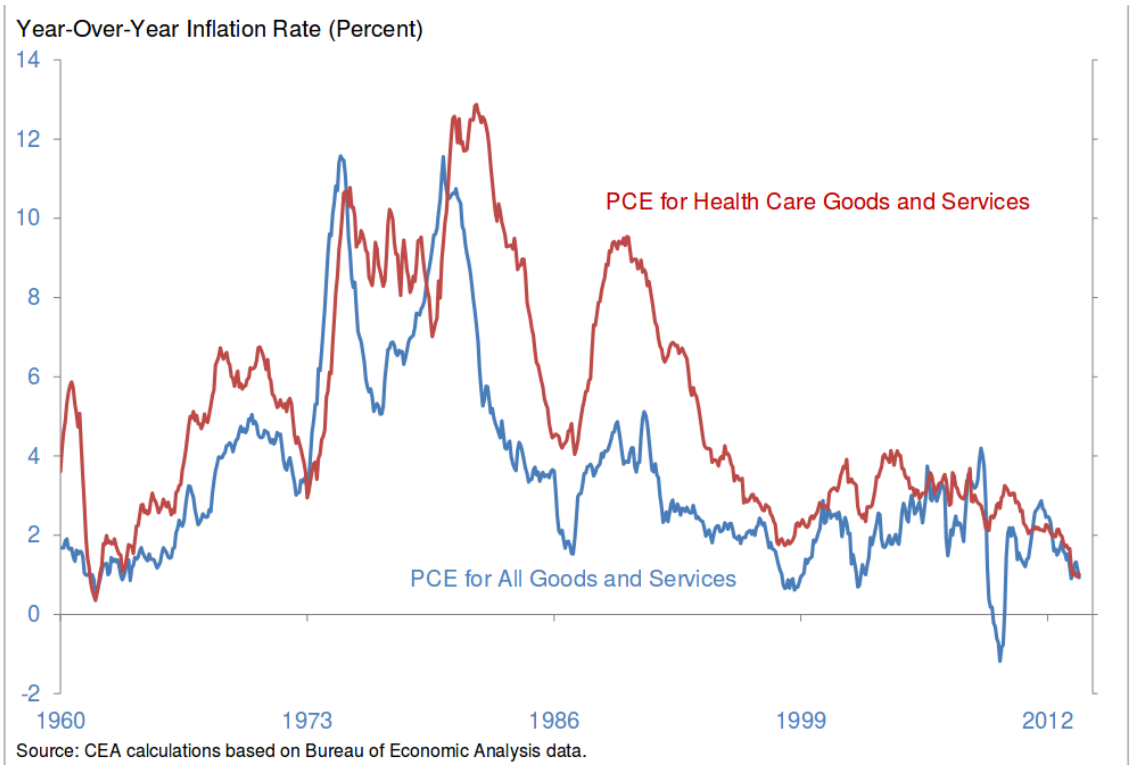


Figure 1.1: Healthcare Inflation versus General Inflation: Historically, general inflation (blue) has been outpaced consistently by healthcare inflation (red). Recently, this has stabilized, but expenditures 17.9% of GDP: the highest in the rich world by far. Source: (Staff, 2013) (Public Domain)

Despite the need for new approaches and ways of thinking about healthcare, a new paradigm has yet to emerge. Research dollars that ought to focus on reducing these costs through prevention, early detection and risk-assessment technologies are instead spent on expensive interventional projects with low success rates that, even if successful may only extend lives by months due to their interventional focus. Despite increased investment between 1998 and 2008, the number of approved new molecular entities (drugs) has declined significantly (Pammolli, Magazzini, & Riccaboni, 2011), and clinical trial attrition rates have increased dramatically (Pammolli et al., 2011) (**Figure 1.2**). Put simply, it is becoming harder to develop new drugs, indicating that a change is needed in which treatments to pursue preclinically. These attrition rates are market

driven, because incentive schemes are geared toward drugs that are profitable. The profitable classes of drugs are interventional, and a more effective drug (one that passes phase III trials) is a more profitable drug. The investment piles on, which drives profits (if at the cost of increased attrition) under the current system.

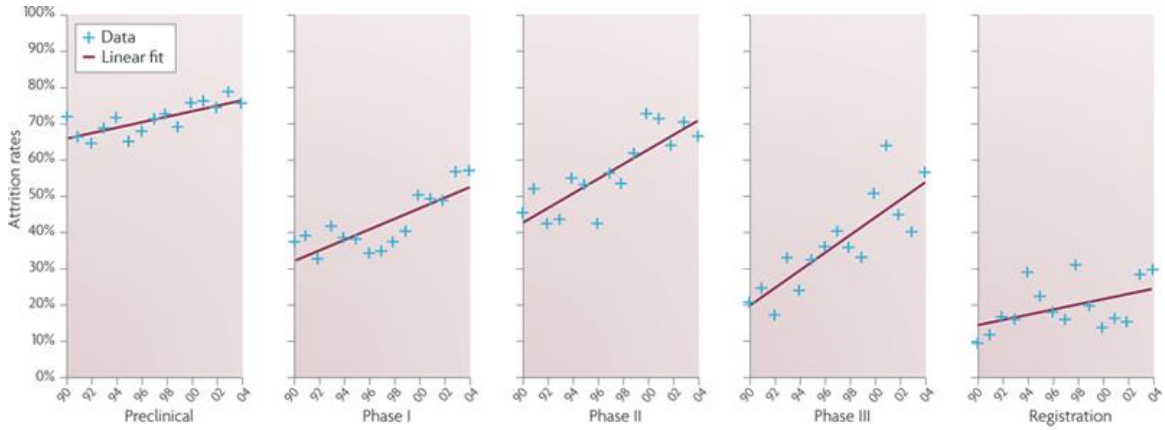


Figure 1.2: Attrition Rates Over Time for All Clinical Trial Phases: There has been a dramatic increase in clinical trial attrition rates since 1990, suggesting that new drug interventions are becoming more difficult and expensive. Reproduced with permission: (Pammolli et al., 2011)

Prevention: Return on Investment

The conventional wisdom on preventative medicine (which includes early diagnostic screening) is that it tends to add to healthcare expenditures rather than decrease them. This is based on hundreds of studies showing that less than 20% of preventative healthcare options are cost-saving, while 80% add cost (Russell, 2009). Though the same studies predict prevention approaches are more effective than intervention approaches in terms of health outcomes, the approach is unviable if prevention increases costs. The system simply cannot afford quality prevention at this time for all but the wealthiest in society.

This conventional wisdom, though accurate, typically draws the flawed conclusion that research into preventative and early diagnostic approaches is not worthy of intense investment. Instead, interventional treatments receive billions in yearly investments from pharmaceutical companies (Mestre-Ferrandiz, Sussex, & Towse, 2012), in addition to the $30.1 billion budget of the NIH (Loscalzo, 2006), which has had almost no impact on the incidence, prevalence or number of days spent in the hospital for *any* disease (Gross, Anderson, & Powe, 1999). Why is the public spending so much for so little?

With this level of investment in interventional treatments, these should be much more effective and cost saving than preventative treatments. Instead, the impact is marginal, the efficiency is getting worse, and healthcare costs continue to spiral unnecessarily out of control, burdening every American healthy or otherwise. While prevention and early screening is at this time is a risky bet, in the author's opinion it is far less risky than the current regime, which has proven to be expensive, ineffective and most beneficial to the small class of wealthy individuals and companies that are able to profit off of desperate people in the late stages of fatal disease.

It is time to give prevention, early diagnostics, and early treatments a serious look. The failures of the past regarding prevention are hurdles that can be overcome. A preventative approach can be cost saving only if it is inexpensive, highly specific, highly sensitive, and approachable to the clinician to encourage wide adoption. Smart allocation of research dollars and an honest acknowledgement of these requirements, if met, can change the situation. Modern biology allows us to take very detailed measurements at annually decreasing costs (Shendure & Ji, 2008).

Through effective analysis and good platforms, these measurements can be turned into valuable information which can be used to affect real reductions in the incidence and prevalence of disease. For the first time in history, cheap, highly sensitive and highly specific diagnostic technologies are within our grasp. This is due to major advances in the tools we use, our understanding of basic biology, and of disease processes. This thesis covers in detail one such technology aimed at this lofty but noble goal, but it is important to note that this is just one of many possibilities in an ever growing ecosystem. In the skilled hands of clinicians who are currently "flying blind" due to an outdated healthcare delivery and research model, modern biology combined with technology could have a revolutionary impact on how the human race approaches disease, diagnostics and treatment.

Biomarkers for Prevention

One such class of modern measurements that could shift healthcare towards prevention and early diagnostics are biomarkers. These are typically (but not limited to) circulating molecules that indicate, correlate with or predict disease. Cholesterol (LDL, HDL) for example is an important biomarker for cardiovascular disease (Sharrett et al., 2001). More recently, a mutation in the *BRCA1* gene has emerged as an important predictor for breast cancer (James, Quinn, Mullan, Johnston, & Harkin, 2007). Most of these molecules, however, must overcome the blood dilution problem. This means that in order for a biomarker to be a useful predictor of disease, there has to be so much of it present that often disease has progressed to a troublesome state where easy intervention is impossible (Hori & Gambhir, 2011). Due to the amount of blood in the body combined with the relatively small systemic effect of early stage fatal disease, if a biomarker is

present in sufficient concentration to be detectable above normal range, something has already gone seriously wrong. What is needed is a class of biomarkers that can amplify signals before symptoms arise, enabling cheaper, more effective diagnostics and early interventions. Though not biomarker based, one of the few examples where this approach is widely practiced is that of colorectal cancer screening. Treatment for stage I or stage II colon tumor (or any in-situ carcinoma) is a relatively straightforward surgery and effectively a cure, but this is only possible if it is caught early enough through regular screening. This is currently not possible for most lethal diseases. Developing the proper biomarker-based diagnostics could change this.

Antibody Biomarkers

Of all the molecules in the human body, there is one class that stands out as an ideal biomarker. Antibodies are generated by the humoral immune system, which is constantly monitoring and reacting changes in the blood and most organ systems of the body. B-cells which produce antibodies go through a rapid process of hypermutation, selection and clonal expansion (Teng & Papavasiliou, 2007) in response to a potentially dangerous foreign entity (antigen). One B-cell clone can rapidly secrete many immunoglobulin molecules (Halperin, 2011), meeting the amplification requirement for overcoming the blood dilution problem. Traditionally these antibodies have been thought of as a response only to infectious diseases and autoimmune diseases, the evidence for which has been reviewed in detail by others (Halperin, 2011; Kukreja, 2012; Navalkar, 2014). However, it is fairly clear at this point that there is an immune response to many different chronic diseases and cancers. Some argue that natural anti-cancer antibodies are generally germline or near-germline IgM molecules which bind to carbohydrates

expressed on cancer cells (Vollmers & Brändlein, 2007). Other, more recent papers claim that natural and specific self-IgG molecules are abundant and ubiquitous in human sera, and respond to chronic diseases such as cancer (Nagele et al., 2013; Stafford et al., 2014). Antigen specific IgG molecules have been isolated in stage IV breast cancer patients (M. H. Hansen, Ostenstad, & Sioud, 2001), prostate cancer patients (Dunphy & McNeel, 2005) and in lung cancer patients (Klotz et al., 1999). There is also some evidence that carcinomas can produce and secrete their own IgG molecules in the absence of any contribution from B-cells or the typical humoral immune system (Chen & Gu, 2007). Despite the many studies showing evidence of specific IgG molecules associated with cancer, the traditional view still holds with many scientists that natural antibodies directed against cancer are primarily nonspecific, near germ-line IgM molecules. Recent data challenges this assumption, and very recent attempts at using these specific "non-traditional" antibodies as tools for diagnostics have shown promise (Anders, 1986; "DENV DetectTM IgM CAPTURE ELISA," 2012; Restrepo et al., 2011a; Scherf, Lopez-Rubio, & Riviere, 2008).

Detecting Antibody Biomarkers

The various methods of detecting circulating antibodies as biomarkers have been thoroughly reviewed by others (Arnon, Tarrab-Hazdai, & Steward, 2000; Halperin, 2011; Navalkar, 2014).These include phage display techniques, which express millions to billions of random peptides displayed on phage particles and use various enrichment rounds to find populations of peptides which bind the molecule of interest with high affinity (Fack et al., 1997; Krumpe et al., 2006; Paschke, 2006; Rodi, Soares, & Makowski, 2002; Wang & Yu, 2004; Yip & Ward, 1999). This is probably the most well

known and widely adopted antibody profiling technology. Another important method uses protein arrays (Ramachandran et al., 2004; Schweitzer, Meng, Mattoon, & Rai, 2010). These typically spot or synthesize recombinant proteins on various surfaces using various methods and apply serum or monoclonal antibodies to the resulting arrays. Using a fluorescently labeled secondary antibody, protein-antibody complexes can be identified. The other major class of antibody profiling technologies is peptide arrays. These arrays either tile known antigens or epitopes in a targeted, designed way (Chen et al., 2010; Forsström et al., 2014), seek to represent a large portion of pathogen proteome space in an attempt at a high throughput multiplexed diagnostic(Navalkar, 2014), or simply spot or synthesize non-natural random sequence peptides en-masse with the hope that antibodies will bind in a reproducible specific way (Halperin, Stafford, & Johnston, 2011; Kukreja, Johnston, & Stafford, 2012; Legutki, Magee, Stafford, & Johnston, 2010; Legutki et al., 2014; Reineke, 2004; Reineke & Sabat, 2009; Restrepo, Stafford, & Johnston, 2012; Restrepo et al., 2011a; Sykes et al., 2013). The latter type of arrays will be introduced and tested in detail throughout this thesis, but first it is important to consider the characteristics and properties of the antibody-epitope interactions, with a special emphasis on continuous (linear) peptide interactions.

Antibody-Epitope Interactions

Antibodies can bind a wide class of molecules, including carbohydrates, small molecules and peptides/proteins. The remainder of this thesis focuses on antibody-peptide and antibody-protein interactions, but it is important to note that antibodies are not limited to these. When an antibody is raised against and binds strongly to a protein, that protein is called the antigen. The region on the protein to which the antibody binds is

the epitope, and the corresponding region on the antibody side is called the paratope. Since proteins are folded in all sorts of ways, the antibody paratope may bind a continuous (linear) portion of the protein or a discontinuous (two or more separate linear sections) portion of the protein. Even if an epitope is continuous, it may require a particular structure supported by the rest of the antigen (constrained continuous epitope), meaning a peptide fragment from the epitope existing outside the context of the rest of the antigen would not bind the paratope.

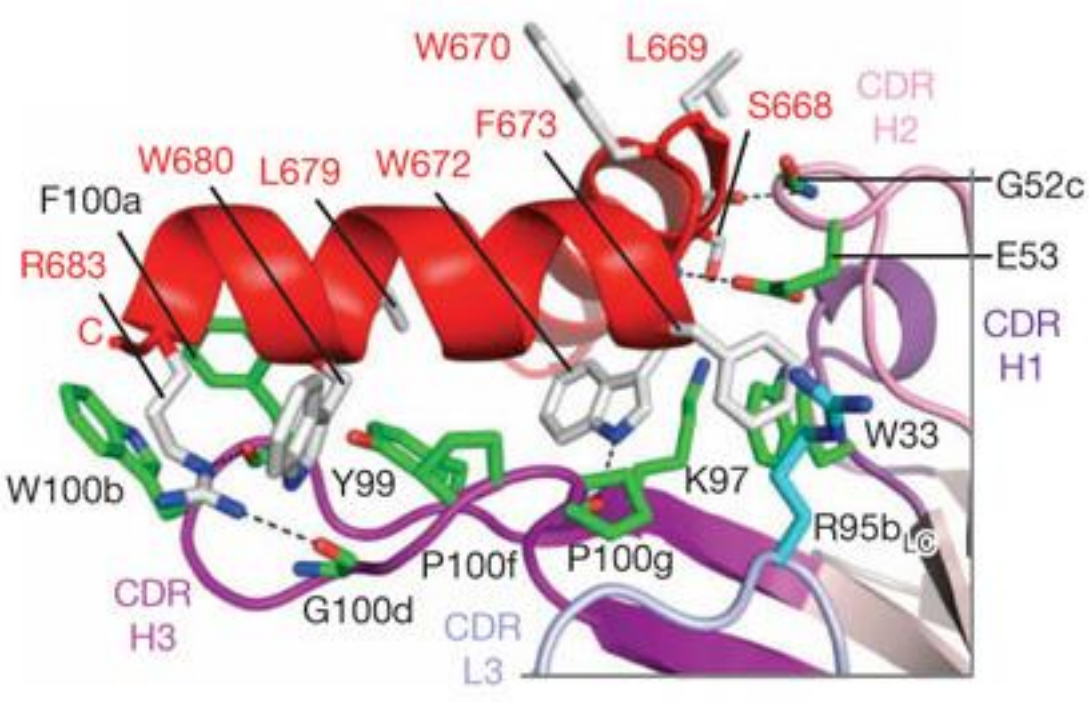


Figure 1.3: Crystal Structure Example of Linear Epitope Bound to Paratope: This is a partial view of the crystal structure for the monoclonal antibody 10E8 (purple, green) in complex with HIV protein gp41 (red, grey). This interaction is an example of a continuous epitope. Note that bound residues from the epitope are linearly close together (W680, L679, W672, etc). Reproduced with permission from (Huang et al., 2012).

Both situations are commonly found in nature, and crystal structures exist of both types (Huang et al., 2012; Karpusas et al., 2001). An example of each is shown in **Figure**

1.3 (continuous) and **Figure 1.4** (discontinuous). The evidence shows that most epitopes are somewhat discontinuous (they don't bind all residues in a row), but many of these residues are spaced closely enough together such that they are effectively linear.

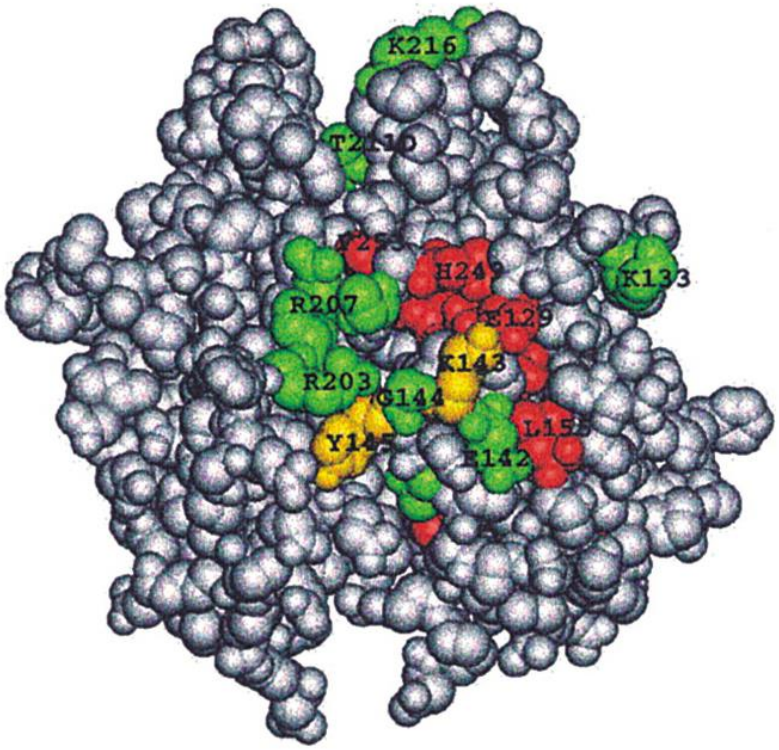


Figure 1.4: Example of a Discontinuous Epitope Space Filling Model: This space filling model is from the crystal structure of CD40 in complex with a neutralizing humanized antibody. Mutational studies of the residues around the binding pocket revealed the residues important to complex formation (red, yellow). Note the labels on the residues, and that that some are hundreds of positions apart (H249, E129). Reproduced with permission from (Karpusas et al., 2001).

The exact ratio between continuous and discontinuous epitopes in nature is unknown, but a survey of the literature on crystal structures (along with data presented in this thesis) shows that continuous, linear and semi-linear epitopes are a very important and common class of epitopes (Huang et al., 2012; Niederfellner et al., 2011; Stegmann, Lührmann, & Wahl, 2010).

Size and Characteristics of Linear Epitope Interactions

If one wants to capture and measure antibodies, one ought to know the types of structures that an antibody binds. To simplify the problem, we limit ourselves to considering only the linear epitopes. Note that in **Figure 1.3** not all the residues in the complex contribute equally to binding. Few studies have been done to determine the number and distribution of residues required for a mature monoclonal antibody paratope to bind with high affinity to its epitope. One good way to do this with linear epitopes is to do a complete substitution study. This is where, for each position s_i on the linear epitope sequence *S*, the amino acid at that position is replaced by every other amino acid, and a fluorescent binding assay is conducted for each "mutated" peptide against the antibody of interest. The ratio of fluorescence between the original sequence and the position specific mutated sequence provides a measure of the contribution to the amino acid at position s_i. More precisely, for each position in the sequence *S*

$$S = s_1, s_2, ..., s_i, ..., s_n$$

and each amino acid a_j from the possible amino acids *A*

$$A = a_1, a_2, ..., a_{20} = [R, H, K, D, E, S, T, N, Q, C, G, P, A, V, I, L, M, F, Y, W]$$

form an n x 20 matrix of peptides *P* where

$$P_{i,j} = s_1, s_2, ..., s_{i-1}, a_j, s_{i+1}, ..., s_n$$

Using a fluorescence assay (ELISA, arrays, others) a measurement is taken that is correlated to binding strength of an antibody with the sequence $P_{i,j}$, forming an n x 20 matrix of fluorescence values *F*. Dividing *F* by the measurement obtained from the fluorescence assay of antibody with the original sequence *S* gives the position specific scoring matrix (PSSM). This measures the relative effect of each amino acid substitution on observed binding. An example of a PSSM in the context of antibody-epitope interactions is given in **Figure 1.5**.

In the Center for Innovations in Medicine this approach has been used to optimize ligands for synbodies (Greving et al., 2010), but has also been used by others to estimate the lengths and number of amino acids contributing to epitope-paritope complex formation. Possibly the most comprehensive study undertaken to answer this question was conducted by Buus et al. in 2012 (Buus et al., 2012b). They raised polyclonal rabbit antibodies against 22 protein fragments (PrESTs) ranging from 50 to 150 amino acids long. They found and defined 49 15-mer epitopes against these anti-PrEST antibodies through exhaustive subsequence searching on arrays, and then did the substitution process described above for each of these 49 epitopes with its target PrEST. Using this process they were able to create a PSSM and identify the important amino acids contributing to epitope-paratope complex formation. After repeating this process for all the PrEST-epitope combinations, they estimated a distribution for epitope lengths. This is reproduced in **Figure 1.6**.

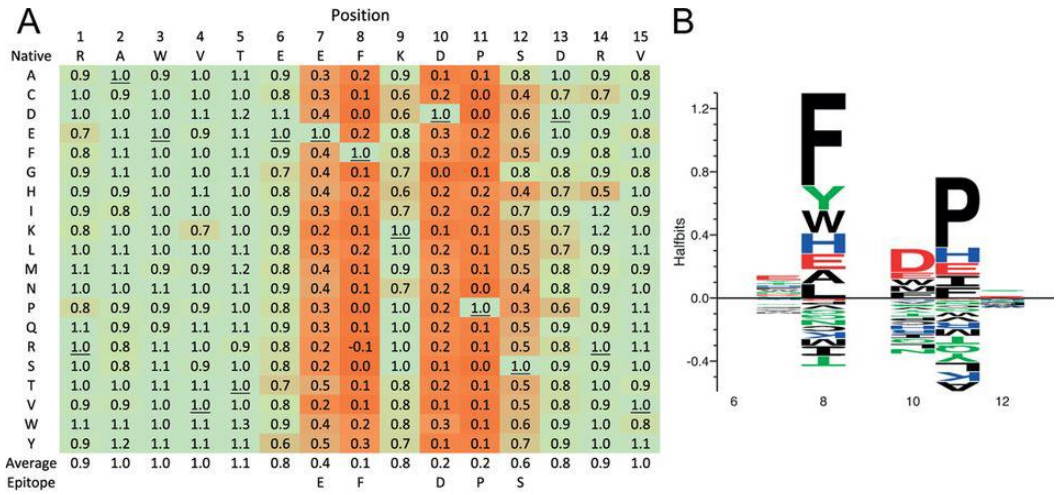


Figure 1.5: An Example of a Position Specific Scoring Matrix (PSSM) For a Polyclonal Antibody Against a Defined Epitope: (A): By doing comprehensive substitutional analysis, important amino acids can be determined for epitope-paratope complexes and the length of the interaction can be estimated with a PSSM. Low values indicate substitutions that inhibited complex formation. (B) The corresponding motif cartoon for the interaction as determined by the PSSM. There are four contributing residues with various allowable substitutions. Reproduced with permission from (Buus et al., 2012b). The figure was adapted from its original form. This is permissible under the Creative Commons 3.0 License http://creativecommons.org/licenses/by/3.0/.

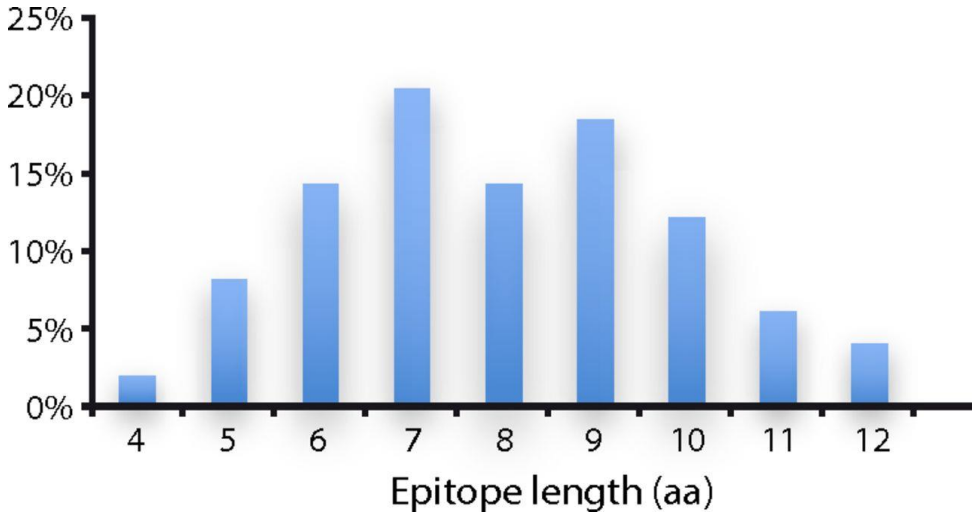


Figure 1.6: Estimate of Linear Epitope Lengths Through Comprehensive Substitutional Analysis: This distribution represents the range of possible interaction lengths for epitope-paritope complexes. Reproduced with permission from (Buus et al.,

2012b). The figure was adapted from its original form. This is permissible under the Creative Commons 3.0 License http://creativecommons.org/licenses/by/3.0/.

A re-analysis of the Buus data revealed that though continuous epitope length is centered on 8 amino acids, there are fewer residues within the total length that are actually important. Depending on the PSSM cutoff, there are on average between 4 and 7 residues contributing to binding in an epitope-paratope complex. These results are summarized in **Figure 1.7**.

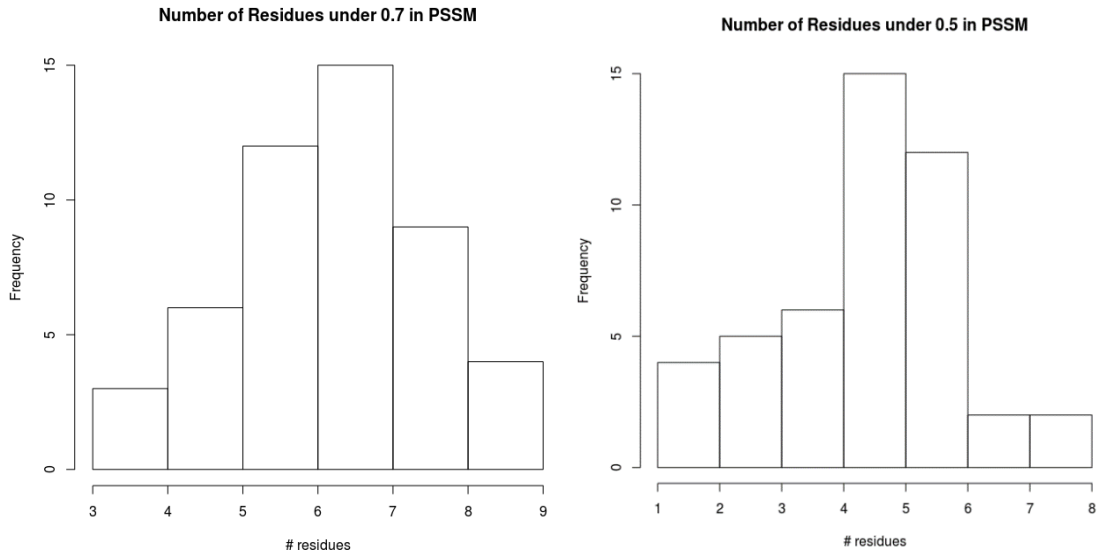


Figure 1.7: Estimate the Number of Important Binding Residues in Epitope-Paratope Interactions: These distributions are based on reanalysis of the Buus data. It concludes that on average, 4 to 7 epitope residues are important to binding in an epitope-paratopecomples. Reproduced with permission from (Buus et al., 2012b). The figure was adapted from data contained within. This is permissible under the Creative Commons 3.0 License http://creativecommons.org/licenses/by/3.0/.

To summarize, epitopes exist in continuous and discontinuous forms. It is not clear what proportion of antibody epitopes is continuous, but it is clear that they are common. These "linear" epitopes consist of a number of binding residues with an average

length covering 8 amino acids, 4 to 7 of which actually contribute to epitope-paratope complex formation.

Mimotope Interactions

In addition to specific epitope-paratope interactions, another type of interaction exists. These are the mimotopes, and they are often uncovered during phage display experiments. These are sequences that bind paratopes with high affinity but do not share sequence similarity with the epitope from which an antibody was raised (Smith & Petrenko, 1997). This explains the common problem of cross reactive monoclonal antibodies, and often confounds epitope mapping experiments. However, these interactions may be useful to immunosignatures because they expand the space to which an antibody can bind. There have been many attempts to use mimotopes discovered in phage display experiments as vaccines (Arnon et al., 2000; Riemer et al., 2005; Wagner et al., 2005). The properties of these interactions (crystal structures, substitution analysis) have not been widely studied, but the few crystal structures that exist suggest that mimotopic peptides bind a paratope in a similar fashion to a true epitope (Saphire et al., 2007). See **Figure 1.8** for a schematic of a mimotope-paratope interaction based on a crystal structure. It is interesting to note that for this particular interaction, the number of residues and length of the interaction is longer than that predicted by the Buus data. Perhaps mimotopes rely on more but weaker residue contacts than true epitopes.

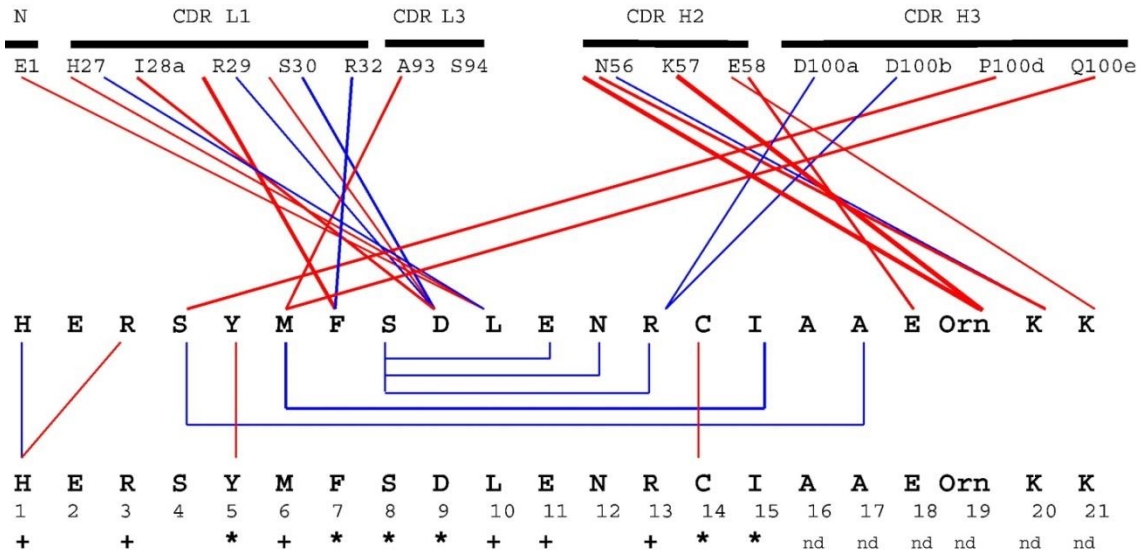


Figure 1.8: Mimotope-Paratope Interaction Schematic Based on Crystal Structure: This schematic shows the residue interactions between an anti-gp120 monoclonal antibody and a mimotope peptide that competes with gp120 for binding. Figure reproduced with permission from (Saphire et al., 2007). Note that this interaction is longer and involves more residues than that predicted by the Buus data.

Random Peptide Arrays and Immunosignatures

Now after a thorough review of antibodies, epitopes, and the details of how they interact, we return to the topic of using peptide arrays as a diagnostic device for early detection and preventative medicine. This approach was pioneered by Dr. Stephen Albert Johnston, Dr. Phillip Stafford and Dr. Neal Woodbury in the Center for Innovations in Medicine at Arizona State University. The basic idea relies on the fact that there are innumerable antibodies circulating in the blood at any given time. These antibody populations are constantly changing and adapting to new threats and circumstances within the body. Though the scope of this surveillance is unclear, given the wide body of evidence reviewed earlier in the chapter, one can assume the antibody repertoire responds in some way to most disease states a human will face.

Assuming this (and it is not a big leap given the evidence already discussed), the problem becomes how to detect these changes robustly and reproducibly. The number of possible linear epitopes is large, but given that the length and number of important residues has a known distribution, it is not infinite. Therefore, if enough peptides could be assayed against the antibodies contained in a person's blood, one could develop a profile, or "immunosignature" of that person's antibody repertoire at any given time. By monitoring this signature over time, or comparing it to a population of signatures from many people, one could associate patterns in the signature with (theoretically) any disease state. The requirements for such a technology are that it be cheap, robust, and highly sensitive and specific so that it could be used for routine monitoring. If these requirements are met, there is hope that the era of expensive and ineffective late-interventional medicine will end.

The solution developed in the Center for Innovations in Medicine for achieving these goals relies on randomized, non-natural sequence peptide arrays. The idea is to immobilize as many short peptides (12 to 17 aa) as possible on a surface, and use this complex surface to profile antibody repertoires. This is a combinatorial approach, and the elegance lies in the fact that the peptide need not map to any known epitope or antigenic protein. The only object of consideration is the signature, or binding profile observed on the array. The process for developing a diagnostic based on this complex surface is simple: probe a population of patients, some with a disease of interest and some without. Using this collected dataset, identify aspects of the immunosignature that predict disease, and use that model to predict correctly new samples. Such a scheme could be adapted to any disease to which the humoral immune system responds.

Over the past 10 years, there has been much progress in this endeavor by individual contributors and by the group as a whole. Dr. Bart Legutki published the first results from immunosignatures, showing that an array of 10,000 non-natural sequence peptides could distinguish vaccinated from unvaccinated mice, and that this effect was amplified over time (Legutki et al., 2010). He followed this up with another study showing that the same arrays can predict whether a vaccine will be effective (Legutki & Johnston, 2013), a tool that would reduce development time for vaccines.

The methods for analyzing these complex high dimensional data are constantly evolving. Muskan Kukreja reviewed machine learning algorithms for achieving maximum accuracy in immunosignature data, and found Naïve Bayes classifiers to be most effective (Kukreja, 2012). This approach was not widely adopted by subsequent studies, with lab members instead gravitating towards SVM based methods. Justin Brown developed methods based on latent variables and hierarchical models common in psychology research (Brown, Stafford, Johnston, & Dinu, 2011), but these also have not been tested on subsequent datasets.

There have also been publications evaluating immunosignatures from the perspective of a clinical diagnostic. Lucas Restrepo established that there is an immunosignature for Alzheimer's disease, and this could be a feasible method for diagnosis (Restrepo et al., 2012; Restrepo et al., 2011a). The largest study to date conducted for immunosignatures as diagnostics was published by Phillip Stafford, showing near perfect classification in a panel of several cancers using similar arrays that Dr. Legutki used in the original publication (Stafford et al., 2014).

Since these efforts new, much larger and more complex arrays have been developed (CIM330K arrays). These contain over 330,000 unique peptide sequences and over 27% of all possible pentamers in triplicate within those sequences (Legutki et al., 2014). This represents the second generation of tools for immunosignatures, and most of the data and experiments presented in this thesis focus on these new arrays. The manufacturing process for these is fundamentally different from the first iteration in that they are synthesized in-situ on the surface of a silicon wafer. This enables mass-manufacturing and the opportunity to do high resolution epitope mapping in a similar manner of Buus discussed earlier. Using these new arrays for epitope mapping is a major component of this thesis.

Sequence Space	Example	Size Calculation	Size	# on Arrays	% covered V7 Long
All 4-mers	AAAA AAAV AVHA YYYY	20^4	$1.6 * 10^5$	$5.9 * 10^4$	36.9
All 5-mers	AAAAA AAAVA AVHAD YYYYY	20^5	$3.2 * 10^6$	$5.2 * 10^5$	16.4
All 6-mers	AAAAAA AAAAAY YYYYYY YYYYYD	20^6	$6.4 * 10^7$	$1.6 * 10^6$	02.6
All 4-aa arrangements within 8-mers	AA.AA... A.V.A.A A...AVY Y.Y..Y.Y	$20^4 * \sum_{i=2}^{i=6} \binom{i}{2}$	$5.6 * 10^6$	$2.06 * 10^6$	36.7
All 5-aa arrangements within 8-mers	AA.AAV.. A.V.AYA A..KAVY Y.YQ.Y.Y	$20^5 * \sum_{i=3}^{i=6} \binom{i}{3}$	$1.12 * 10^8$	$1.82 * 10^7$	16.2
All 6-aa arrangements within 8-mers	AA.AAVY. A.VKAYA A..KAVYQ Y.YQ.YKY	$20^6 * \sum_{i=4}^{i=6} \binom{i}{4}$	$1.34 * 10^9$	$3.29 * 10^7$	02.5

Figure 1.9: Definitions and Sizes of Various Epitope Spaces: Linear epitopes can be defined in various ways. Each of these "spaces" has a different size and different characteristics. Some of these spaces are covered very well by the arrays used in this thesis, others are not. In subsequent chapters, these definitions will be tested and used to identify epitopes related to disease, characterize array results, and develop software to rapidly databases for motifs within these spaces.

Project Overview

This thesis explores, characterizes and establishes the limits of the new generation of immunosignature technology. It builds on the work others have done with the smaller arrays, and expands this in terms of the new ones, while incorporating sequence information in new ways. Previously, individual peptides were considered simply as "black box" features, with no regard to the content of those sequences. This thesis changes that, by elucidating the role of sequences in immunosignatures. The first

question, addressed in chapter 1, is whether the new arrays have sufficient complexity needed to map epitopes in monoclonal antibodies and sera. Depending on how one defines an epitope, these arrays contain a high proportion of all possible epitopes, so I develop methods to resolve these from the complex data. **Figure 1.8** summarizes the coverage of the various "sequence spaces" by version 7 of the CIM330K arrays.

The next chapter concerns itself with developing a diagnostic for Dengue Fever. This is evaluated against existing methods, revealing distinct advantages to the immunosignature approach. In the course of this development, a robust classification methodology and software based on SVM was developed and made available under the BSD License.

The third chapter develops a kinetic theory for how multiple antibodies bind multiple array peptides. From this, a testable hypothesis is derived and tested against existing data. This model represents the best guess to the underlying kinetic properties of the assay, and experiments are suggested to further validate and refine the model.

The fourth chapter is a meta analysis considering amino acid content and how well the patterns observed on the arrays are explained by a simple compositional linear model. There is remarkable stability across many different types of arrays and experimental conditions. There is a strong role for charged residues in determining the overall immunosignature pattern, and this effect seems to have some predictive ability for disease.

The final chapter develops a novel indexing strategy and software for rapidly searching through large sequence databases for linear epitope-like patterns. This was developed for a specific project whereby an unknown pathogen should be identified using

only sequence data gleaned from array experiments. Such an approach required rapid pattern searching technology that did not yet exist. Though the original project for which the software was designed did not yield the expected results, the software itself is the fastest known search method and indexing strategy for certain types of patterns.

The overarching theme in this thesis is that sequence information provides unprecedented detail and a new dimension to immunosignatures which still needs further exploration. I lay the groundwork and uncover some important results, particularly with regards to epitope mapping. However, there is always more work to do, and it is my hope that the work laid out here will be yet another pebble upon which others can build.

CHAPTER 2

EPITOPE IDENTIFICATION USING PEPTIDE MICROARRAYS

Abstract:

Antibodies play an important role in modern science and medicine. They are essential in many biological assays, and have emerged as an important class of therapeutics. Unfortunately, current methods for mapping antibody epitopes require costly synthesis or enrichment steps, and no low cost universal platform exists. In order to address this, we tested a random sequence peptide microarray consisting of over 330,000 unique peptides sequences sampling 83% of all possible tetramers and 27% of pentamers. It is a single, unbiased platform capable of performing many different types of tests, it does not rely on informatic selection of peptides for a particular proteome(s), and it does not require iterative rounds of selection.

In order to optimize the platform, we developed an algorithm that considers the significance of k-length peptide subsequences (k-mers) within selected peptides that come from the microarray. We tested eight monoclonal antibodies and seven infectious disease cohorts. The method correctly identified 5/8 monoclonal epitopes, and identified both reported and unreported epitope candidates in the infectious disease cohorts. This algorithm may greatly enhance the utility of random-sequence peptide microarrays, by enabling rapid epitope mapping and antigen identification.

Introduction:

Antibodies play a central role in the immune system and in modern healthcare and medical research. They are commonly used as affinity reagents in research and diagnostic applications, and have emerged as an important class of therapeutics (Stafford et al.,

2014). When generating new affinity reagents, it is useful to know the target sequence (epitope) bound by the antibody in question. Many methods have been developed to accomplish this, including peptide tiling and phage, bacteria and mRNA display (Ballew et al., 2013; Fack et al., 1997; Reineke, 2004). In disease diagnosis, especially for newly discovered diseases like MERS (Zaki, van Boheemen, Bestebroer, Osterhaus, & Fouchier, 2012), knowing the epitope(s) that elicits a humoral response enables production of diagnostics and vaccines. Large-scale mapping of cohorts infected with the same disease may guide development of universal vaccines for flu and other infections. Crystal structures provide the most information about antibody–antigen binding for linear or conformation epitopes, but in practice this is cost prohibitive and rarely done. Display or library panning-type approaches use bacteria or phage to display peptide sequences, avoiding costly crystallization or synthesis steps and area common approach for linear epitope mapping (Fack et al., 1997; Paschke, 2006). Recently, bacterial display methods have been used to discover antigens to Celiac disease (Ballew et al., 2013). Tools for probing the "memory" of the immune system could reveal a wealth of information about an individual's health status and antibody repertoire. While display techniques are effective and result in highly accurate and specific linear epitope determination (Wang & Yu, 2004; Yip & Ward, 1999), they have hidden and poorly understood biases regarding sequence populations (Krumpe et al., 2006; Luck & Travé, 2011; Rodi et al., 2002) and rely on selection steps that eliminate certain sequences in favor of others. This process creates issues with cost and reliability at scale, and information is discarded as the selection process becomes increasingly stringent. As a rapid identification method, panning is non-optimal.

Peptide array technologies provide an alternative approach. They are simple, reproducible and low cost if mass produced, but represent a smaller sequence library than phage display and only contain linear sequences. This is an apparent disadvantage, but in practice, linear epitopes are actually quite common in nature and even mimotopes could provide useful, if indirect, information about non-linear epitopes. Microarrays of random-sequence peptides are amenable to antibodies that strongly and specifically bind peptides consisting of short, gapped sequences containing four to six anchor residues, which seem to cover a sizable class of antibodies (Buus et al., 2012a; Sivalingam & Shepherd, 2012). To date the most common approach to designing peptide microarrays has been to tile sequences from a known protein or proteome of interest, and find sequences that bind the target (Edfors et al., 2014; Forsström et al., 2014; Hansen, Buus, & Schafer-Nielsen, 2013; Reineke, 2004; Reineke & Sabat, 2009). Recently this technique has been scaled to whole proteomes using arrays containing millions of sequences (Forsström et al., 2014; Hansen et al., 2013). This approach is effective on a single protein scale, but problems arise when looking for specific epitope sequences in the presence of millions of other peptides. Cross-reactivity of antibodies to non-target peptides often obscures the eliciting antigen (Forsström et al., 2014). This may be due in part to the fact that tiled peptides are fundamentally different from folded proteins, and inaccessible parts of a protein are likely to be exposed when linear pieces of it are tiled. Additionally, there are many common n-mers across apparently unrelated pathogens (Halperin, 2011; Navalkar, 2014). It may be possible to address this problem using motif-based discovery rather than peptide-based discovery. Brief motifs (4-5mers) should appear redundantly in a given peptide library. Longer (6-12mers) exact sequences should appear more rarely. There

may be some way to leverage the higher confidence in short but redundant motifs in lieu of long but rarer sequences. A platform is needed that focuses on representing as many sequences as possible on an array, rather than accumulating sequences from a particular set of proteins which may or may not mimic the folded structure.

Previously our group has used random-sequence peptide microarrays to diagnose disease using immunosignatures (Stafford et al., 2012; Sykes et al., 2013). The effect relies on the interaction of serum antibodies with random-sequence peptides bound to a microarray to provide information about a disease state (Hughes et al., 2012; Legutki et al., 2010; Restrepo et al., 2012; Restrepo et al., 2011b; Sykes et al., 2013). While immunosignatures are sensitive and specific as a diagnostic, there has not been a link established between the immunosignature profile consisting of peptides differentially bound between healthy and disease cohorts, and the actual sequences of those peptides. This was attempted in a previous study from our group which evaluated an array of 10,000 17-mer peptides as a platform for epitope mapping. We found that while useful for predicting linear sequences to some monoclonal antibodies, it offered virtually no predictive power in serum samples from mice immunized to a known antigen (Halperin et al., 2011). Since then, advances in *in-situ* synthesis techniques have produced arrays containing over 330,000 peptides with the possibility of scaling to several million peptides per array (Legutki et al., 2014). These new arrays contain over 27% of possible pentamers and 83% of possible tetramers. While still lacking the majority of pentamers, this is a fairly dense sampling of short peptide sequences that may be useful for epitope mapping.

Here we report on a general approach to using random sequence peptide arrays to map epitopes. We demonstrate this by recovering eliciting sequences in a set of monoclonal antibodies, and then apply it to a set of disease sera containing antibodies of unknown specificity, revealing both previously discovered and new epitopes. The study described here is the first attempt at deciphering a microarray with fixed but random peptide sequences for epitopes that does not *a priori* assume a set of eliciting proteins.

Methods

Array Construction

Peptide microarrays are manufactured using *in-situ* synthesis of 330,000 random-sequence peptides per 0.5cm^2 region. Each 75mmx25mm slide contains24 sub-arrays, each containing 330,000 peptides. The average length of each peptide is 11.2 amino acids with a standard deviation of ±1.3, normally distributed. The longest peptide is 22aa long, the shortest is 1aa, with 95% of peptides between 8aa and 14aa.Peptides are synthesized from C-terminal to N-terminal, with the amine group furthest from the array surface. Prior to assay, they are washed in 100% DMF for one hour, then introduced to PBST incubation buffer (3% BSA in Phosphate Buffered Saline, 0.05% Tween 20) over a period of six hours to allow the solvent phase to be completely transitioned to aqueous phase. The arrays are then processed by incubating in the presence of antibodies or serum and detected by fluorescent antibody (see methods in (Legutki et al., 2014)).

Binding of antibodies to the array

Residual DMF was removed by two 5 min. washes in distilled water. Arrays were equilibrated in PBS for 30 min and blocked in incubation buffer (3% BSA in Phosphate Buffered Saline, 0.05% Tween 20 (PBST). Arrays were washed and briefly spun dry

prior to loading into the multi-well gasket (Array-It, Santa Clara, CA). Incubation buffer was added to each well (100ul) and 100 ul of 1:2500 diluted sera was added for a final concentration of 1:5000. Arrays were incubated for 1hr at 23ºC with rocking, and then washed with PBST and 1% BSA in PBST using a BioTek 405TS plate washer. Anti-human IgG-DyLight 549(KPL, Gaithersburg, MD) was added to a final concentration of 5.0 nM. Unbound secondary was then removed by washing in PBST followed by washing in distilled water. The arrays were removed from the gasket while submerged, dunked in isopropanol and centrifuged dry (800xG, 5 min).Arrays were scanned at 533nm using an Innoscan 910 array scanner (Innopsys, Carbonne, France). Features were aligned and extracted using GenePix Pro 6.0 (Molecular Devices, Sunnyvale, CA).

Monoclonal Antibodies

Eight monoclonal antibodies were used in this study. Anti-human HA (Rockland Antibodies, Rockland, MD, [YPYDVPDYA]), DM1A (anti-human tubulin, Invitrogen/Life Technologies, [AALEKDYEEVGV]), Ab1 (anti-human TP53 antibodies, Clonetech, Palo Alto, CA, [TFRHSVVV]), FLAG (Invitrogen, Madison, WI, [DYKDDDDK]), 4C1 (anti-human TSHR, Santa Cruz Biotech, Dallas, TX, [QAFDSHY]), A10 (Acris Antibodies GmbH, Hiddenhausen, Germany, [EEDFRV]), Ab8 (Anti-human P53, Thermo Fisher Scientific, Waltham, MA, [TFSDLWKLLPE]), and 2C11 (Acris Antibodies GmbH, Hiddenhausen, Germany, [NAHYYVFFEEQE]).

Serum Samples

Sera from seven different disease cohorts were provided by Seracare Life Sciences (Milford, MA) as well as 10 pools of healthy persons designated as HNP, plus one group of 32 different non-infected sera samples collected from consented volunteers

by the Center for Innovations in Medicine at Arizona State University under IRB# 0905004024 (renewed April, 2014) were used for this study. There were 32 healthy donors (Normals), 9 Dengue Fever (DEN1 Flaviviridae), 8 Lyme disease (*Borrelia burgdorferii*), 7 Syphillis (*Treponema palladium*), 13 Malaria (*Plasmodium falciparum*), 12 Whooping cough (*Bordetella pertussis*), 15 Hepatitis B virus (Hepadnavirus), and 10 mixed pools of normals called HNP (Healthy Normal Pool) that comprised the 8 cohorts.

Analytical methods

Finding Antibody Specific Peptides

The goal of this study is to find sequence motifs corresponding to an epitope. Thus, the first step of an analysis is to identify peptides that bind specifically to the sample of interest without regard to the peptide sequence. First, arrays are normalized to the median intensity value to account for small differences in serum or dye concentrations. Then, fold-change is calculated per peptide across the sample of interest and the median of control samples. In the case of serum samples, the controls were a cohort of non-diseased patient samples. In the case of monoclonal antibodies, the controls were a mix of all other monoclonal antibodies being examined. Based on these transformed values, the top 500 peptides were used as seed sequences for epitope discovery.

Maximal Subsequence Algorithm

The algorithm used to find high binding subsequences was designed to find short consensus motifs within a large set of random peptides. It can be divided into two parts: motif identification and significance testing. Seed sequences are computationally divided into all possible subsequences within a certain range of lengths (3 to 7 amino acids).The

set of these subsequences $\mathbf{S_x}$ are ranked and evaluated for significance in subsequent steps.

The input to the algorithm is a set of sequences $\mathbf{S} = \{s_1 , s_2 , ..., s_n \}$, and associated preprocessed array intensity values $\mathbf{Q} = \{q_1, q_2, \ldots, q_n\}$. To find a set of significant subsequences, the sequences in **S** are divided into all possible subsequences containing between three and seven amino acids each. For example, the sequence AVHAD would be divided into the set {AVH, VHA, HAD, AVHA, VHAD, AVHAD}. All the subsequences in **S** constitute a new set **S'**. Members in **S'** have one or more associated values in **Q** corresponding to the intensities from parent sequences containing that subsequence. We define the function Q_{sub}: $\mathbf{S' \rightarrow Q^m}$, where m is the number of peptides excepting the top 500 seed peptides containing the input subsequence. This gives all intensity values associated with a subsequence.

Sequences $s_i \in \mathbf{S_x}$ are ranked according to their associated values $t_i = Q_{sub}(s_i)$. A subsequence is only considered if it appears in at least three peptides ($|t_i| > 3$). We term this value the *support* of the subsequence. The ranking function considers the support and the median intensity value *median*(t_i), such that the highest ranked subsequences have at least three appearances on the array and have high median intensities . This criterion is not strictly necessary, but it simplifies significance testing by throwing out non-significant, poorly represented sequences. Once subsequences are filtered and ranked, their significance can be established. This occurs for a given subsequence i using the following nonparametric procedure:

1. Draw $|t_i|$ values from **Q** at random. Call this vector t'_i.
2. Compute *median*(t'_i).

3. Repeat steps one and two 10,000 times, resulting in a nonparametric estimate of a t_i null distribution. Call this vector **D**.
4. A p value is computed for subsequence s_i according to: $p_i = \frac{\sum_{k \in D} I(median(t'_i) > k)}{|D|}$ where I is the indicator function.
5. Correct the p values for multiple hypotheses. This work used the following correction function: $p'_i = \frac{p_i}{\sum_{s_i \in S_x} |Q_{sub}(s_i)|}$. For example, if 1000 subsequences are considered, alpha is $\frac{1}{1000}$, resulting in one expected false positive.

Calling Epitope Candidates

Significant subsequences were identified for each individual per disease cohort. In order to determine the most likely epitope candidates, sequences were ranked in terms of the number of subjects in which they were called significant. The sequences that appeared most often in different individuals within the same group were deemed the most likely epitope candidates (**Figure 2.4A**).

Mapping Epitope Candidates to Pathogen Proteomes

The most common significant subsequences (query sequences) were searched against the pathogen proteome for 100% identity. The probability of a match was assessed by searching randomly drawn array sequences of the same length as the query sequence against the proteome, and comparing the expected number of matches to those observed with the query.

Pathogen Identification

Our objective was to identify an unknown pathogen based on array sequence information alone. The (n) significant subsequences from the same cohort were pairwise aligned using the BLOSUM62 substitution matrix, producing an (n x n) matrix of alignment scores. This matrix was hierarchically clustered by single linkage, producing a dendrogram of related subsequences. This analysis revealed peaks of central

subsequences which were presumed to be most closely related to the true epitope. These peak sequences were searched against a database of 596 proteomes (hereafter called the Pathogen Proteome Database) from various strains pathogenic bacteria, viruses, and protists causing over 100 different diseases. Those proteins and organisms matching all queried sequences within 100% or 80% identity were noted. Probabilities were determined by querying the database with randomly drawn sequences as above.

Minimum Required Sequence Information

In order to find the point at which pathogen proteins could be resolved from a database given fixed epitope information, we generated several sets of random sequences ranging in lengths from four to seven amino acids. Pairs of sequences with set lengths were drawn from this set, and queried against two databases: one containing 596 human pathogens, and another containing over 5000 Bacteria, Viruses, and Eukaryotes, in order to determine the point at which pathogens could be uniquely resolved. For example, a trimer sequence would be present in most pathogen proteins, but two heptamer sequences are much less likely to appear in multiple pathogens by chance due to more available sequence information.

Sequence Logo Generation

Significant subsequences were collected together into a FASTA-formatted list. Multiple alignments were produced with ClustalW2 (Chaddock et al., 1995). Multiple alignment text file was used as input to WebLogo 3 (Bähler & Rhoads, 2002) using default settings, producing the motif figure.

E Value Calculations

The reported E-values were calculated by searching random re-orderings (with replacement) of the candidate subsequence against the target proteome, using the mean number of occurrences of 10,000 re-orderings as the E-value.

Results:

We first asked whether we could predicatively map epitopes to well-characterized monoclonal antibodies. Eight antibodies with reactivity to a known linear sequence were chosen and analyzed.

Epitope Determination in Monoclonal Antibodies

Epitope	Ab Name	*Immunogen*	Isotype	pI	GRAVY	Mean Signal Intensity	Mapped Predicatively
EEDFRV	A10	*Human Pol II*	IgG2b	4.1	-1.3	4911	No
SDLWKL	p53ab8	*Human p53*	IgG2b, IgG2a	5.6	-0.3	6243	No
QAFDSH	4C1	*Human Insulin Receptor*	IgG2a	5.1	-1.1	971	**Yes**
RHSVV	p53ab1	*Human p53*	IgG1	9.8	0	5074	**Yes**
DYKDDDDK	FLAG	*FLAG peptide*	IgG1	4	-3.3	1167	**Yes**
AALEKD	DM1A	*Human Tubulin - alpha*	IgG1kappa	4.7	-0.6	5798	**Yes**
YPYDVPDYA	HA	*HA peptide*	IgG1	3.6	-0.9	905	**Yes**
NAHYYVFFEEQ	2C11	*Human Insulin Receptor*	IgG1	4.5	-1	827	No

Table 2.1: Table of Monoclonal Antibodies Used in this Study -- Monoclonal antibodies used to test subsequence analysis. IgG2 was less likely to reveal an epitope than the IgG1 mAbs. Predictive mapping means that the top subsequences were both related to the true epitope and related to each other to an extent that a clear, dominant motif emerged with strong association to the epitope sequence. GRAVY stands form grand average of hydropathicity index(Legutki et al., 2014).

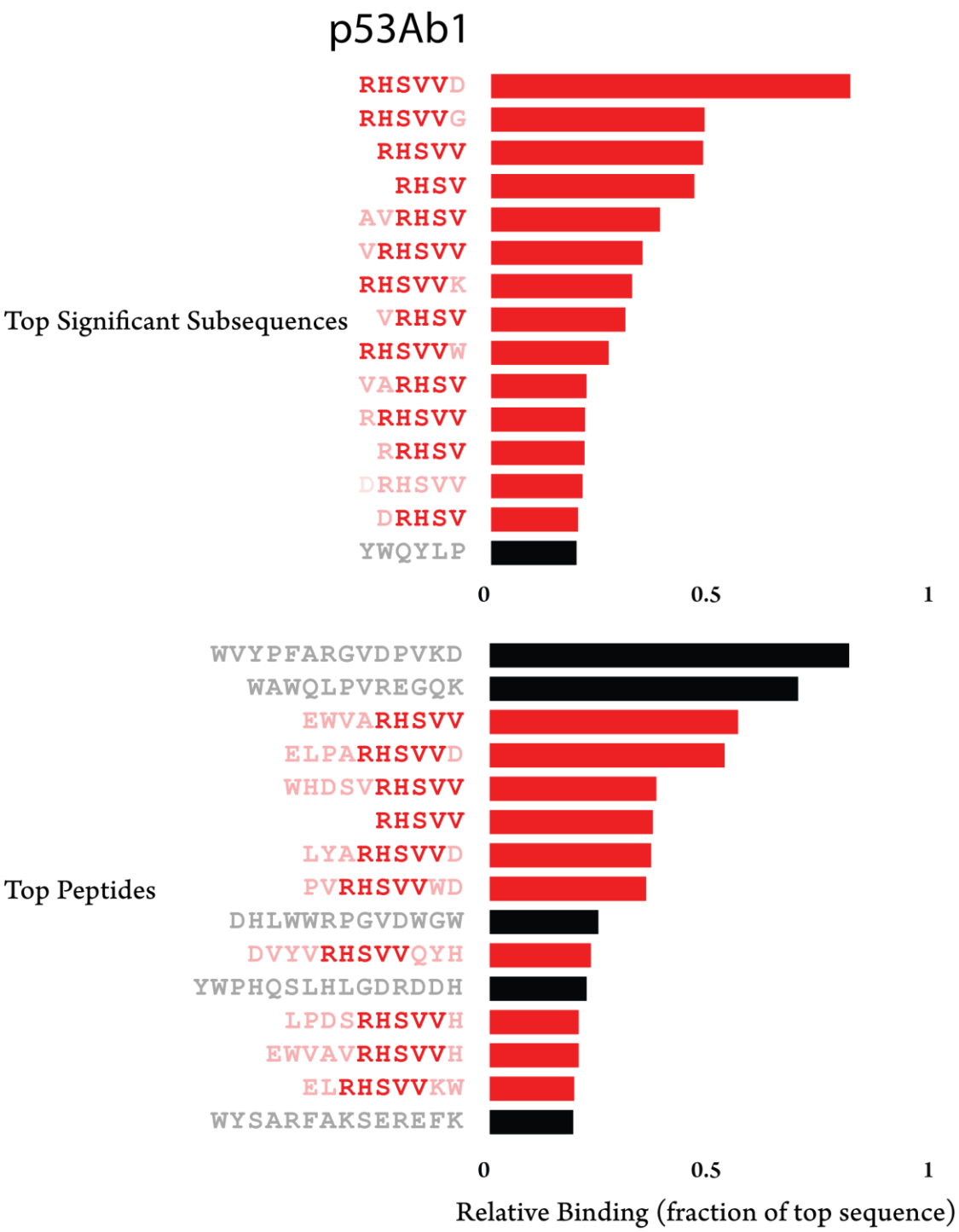
Table 2.1 lists peptides and binding intensities for the 8 different monoclonal antibodies. The linear epitope for each monoclonal is known, and was used as the basis for algorithm development and testing. In most cases, simply sorting peptides by intensity per monoclonal was insufficient to reveal epitope motifs amongst the highest binding peptides. Variation in binding to a specific target comes in part from the amount of non-cognate binding. Highly promiscuous antibodies like anti-HA bind large numbers of peptides with low similarity to the target, creating a lack of specificity in these datasets **(Table 2.2)**.

	Total Binders	On Target	Fraction
AB1	42386	466	1.10×10^{-02}
HA	1608	53	3.30×10^{-02}
4C1	2561	276	1.08×10^{-01}
FLAG	7563	0	0
DM1A	44821	207	4.62×10^{-03}
A10	44924	37	8.24×10^{-04}
AB8	46327	1	2.16×10^{-05}
2C11	671	0	0

Table 2.2: On Target versus Off Target Binding – This table shows the number of peptides for each antibody that yielded a signal greater than 5-fold above background (Total Binders), and of those how many had at least 80% sequence identity with the true epitope (On Target). See Table 2.1 for list of true epitopes. A very low percentage (<11%) of the binding peptides had strong sequence similarity with the true epitope, in agreement with previous studies (Halperin et al., 2011).

However, transforming the data in terms of peptide subsequences revealed highly specific and consistent sequences that corresponded to epitope targets in five of the tested antibodies. Motifs were similar to the exact eliciting peptide sequence. Even when the exact sequence was not present on the array, sequences very similar to the eliciting peptide predominated (**Figure 2.1, Figure 2.2**). Three of the tested antibodies did not generate a specific response to the expected target sequence. In one of these cases

(P53Ab8), the epitope SDLWKL was bound, but due to the high degree of cross reactivity to non sequence-similar peptides, one would not expect to map the epitope based on these results alone (**Figure 2.3A**).



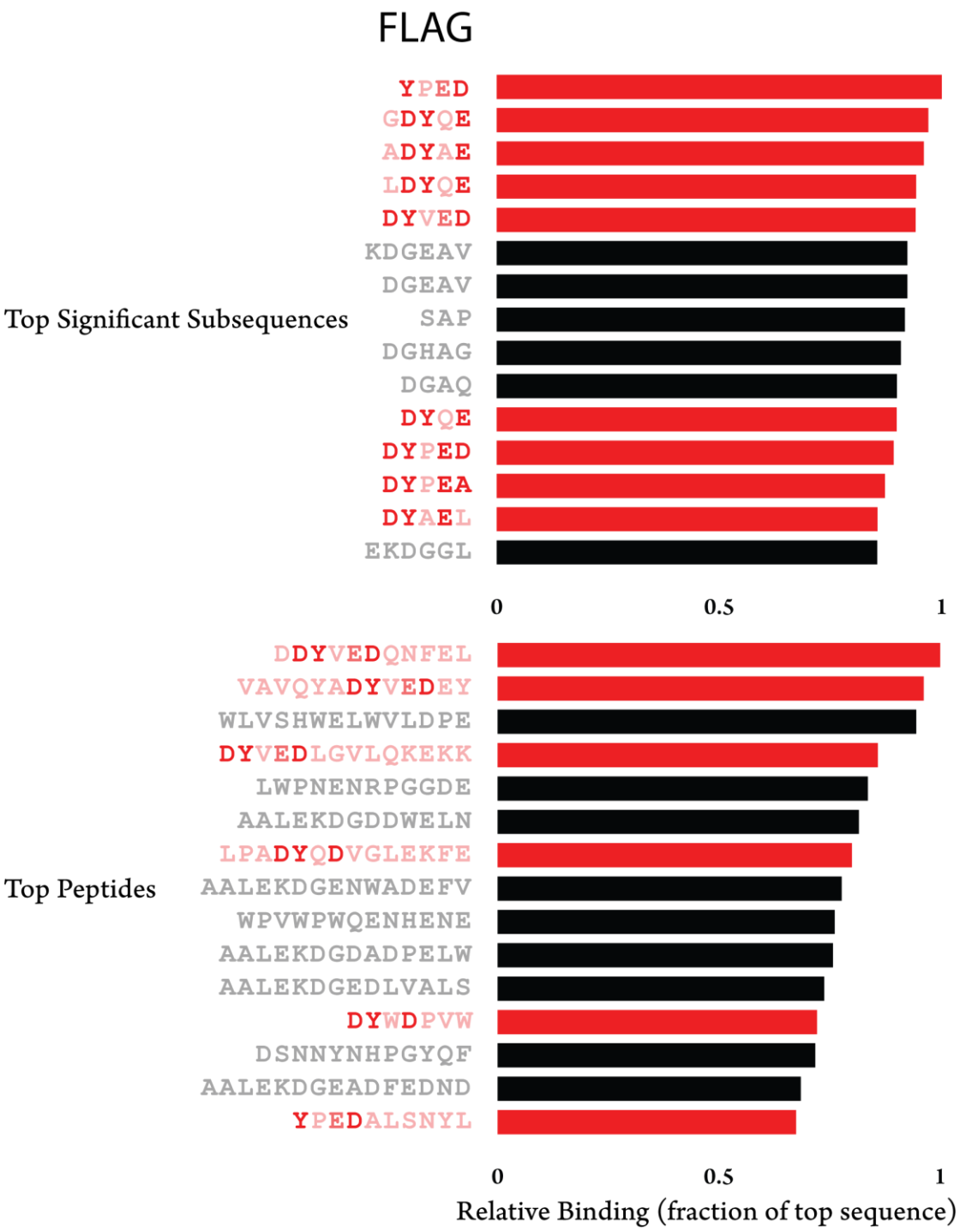
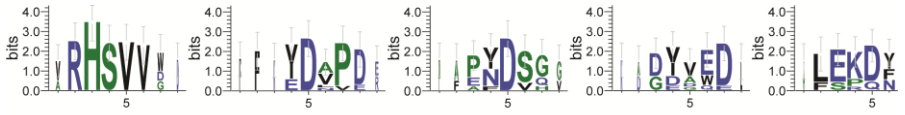


Figure 2.1: Top Binding Subsequences and Peptides for Selected Monoclonals: These bar plots show the top binding subsequences (top) and subsequences (bottom) for two of eight tested monoclonal antibodies. P53Ab1 (RHSVV) and FLAG (DYDDDK) each had on-target motifs throughout the top peptides, and these were made clearer through subsequence analysis (shown in red).

(A)

mAb Name	p53Ab1	HA	4C1	FLAG	DM1A
Motif Cartoon					
True epitope	RHSVV	YDVPD	FDSH	DYDDDDK	AALEKDY
Inter-alignment	23.6	17.4	26.7	16.7	15.9
N subseq.	100	419	678	184	67
Fold Change	13.7 ± 2.7	35.4 ± 3.8	25.9 ± 6.6	45.1 ± 20.55	4.7 ± 0.53

(B)

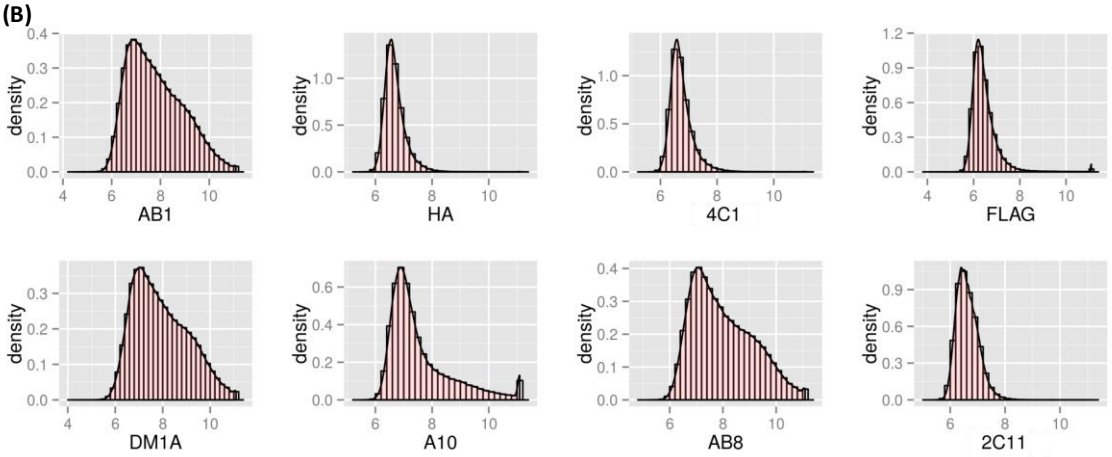


Figure 2.2: Monoclonal Antibody Motifs and their Corresponding Epitopes: (A): These five motifs were revealed after incubating monoclonal antibodies on the peptide microarrays and performing subsequence analysis on the resulting dataset. Sequence logos were created using the top 10 most highly ranked subsequences. Note the positional dependency of the sequences, showing anchor residues as well as regions of apparently low importance to antibody binding. The true epitopes were determined by the manufacturer, and inter-alignment is the expected value of pairwise gapless alignment scores (BLOSUM62 matrix) between any two significant subsequences pulled from the arrays. Fold change indicates the relative binding strength of the peptides making up the cartoon versus the median value for that peptide in the other monoclonals tested. Antibodies for which consensus motifs could not be found were A10 (EEDFRV), p53Ab8 (SDLWKL) and 2C11 (NAHYYVFFEEQE). Additional information about these antibodies and their immunogens can be found in Table 1. (B): Histograms of each monoclonal antibody tested, where the x-axis is log normalized signal intensity. Antibodies varied in their binding profile, with some such as HA, 4C1 and FLAG showing a narrow distribution around low intensities, while others bound many peptides. See Table 2 for an analysis of on target versus off target binding.

The success rate in mapping linear epitopes on monoclonal antibodies was encouraging in that it implied the possibility of quickly mapping disease associated epitopes in patient sera. In order to test this hypothesis, we next performed similar experiments using sera from patients infected with various diseases.

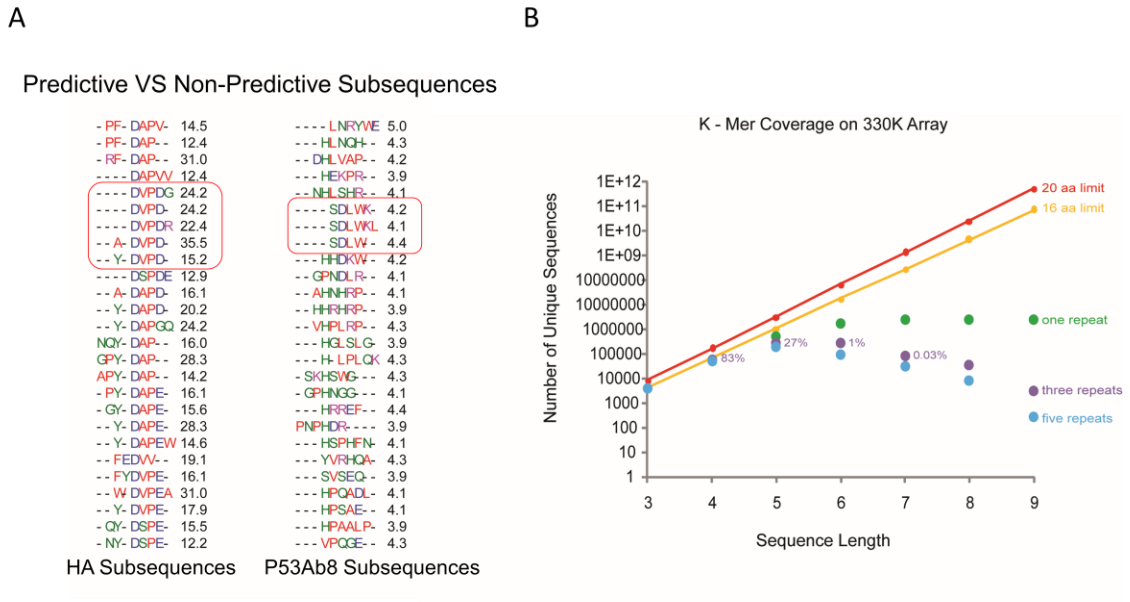


Figure 2.3: Sequence Representation and Predictive versus Nonpredictive Subsequences: (A) In order for the arrays to predict an epitope motif, two conditions must be met. First, the array must sufficiently bind a collection of true epitope or near epitope subsequences. Second, the bound sequences must be stronger than any cross reactions observed on the array. The sequences most related to the true epitope (Table 2.3) are circled in red. The left figure shows the top 25 subsequences and fold changes for the HA monoclonal antibody. While small differences occur, the sequences share a consensus pattern. In contrast, p53Ab1 (right) showed clear binding to the true epitope but cross-reacted with other sequence clusters, preventing good prediction. This result may still be informative when assessing antibodies in a drug or affinity reagent pipeline when specific antibodies are needed. **(B)** The fraction of all possible k-mers present on the array as a function of k-mer length. The arrays represent 27% of all possible 5-mers in at least 3 different peptides. A higher coverage of sequence space should result in better epitope prediction and more comprehensive analysis of cross reactivity.

Groupwise Epitope Determination in Patient Sera

Eight cohorts representing seven different diseases and one group of healthy volunteers, each consisting of between 7 and 15 individual samples, were tested using the \described methods. Several of the cohorts performed similarly to the monoclonal antibodies in that they identified a relatively small number of peptides with highly homogenous sequence motifs that were obvious by simple text matching. These cohorts produced a noticeably homogenous list of peptide sequences that deviated little from a single motif. The multiple alignments of the top 10 sequences for each of these disease cohorts are shown in Figure 2.4B. Of the seven disease cohorts tested, five of these revealed a consensus sequence.

Consensus sequences in pathogen proteomes

In order to test whether the groupwise consensus motifs (**Figure 2.3)** corresponded with true epitopes, we searched the immune epitope database (http://www.iedb.org) for exact substring matches to sequences from our lists. Despite the small size of this database, the sequence AVHAD from Dengue was present in the database and indicated as an epitope from the NS1 protein in two Dengue strains (E-value: $5x10^{-4}$). Further analysis of the other cohorts revealed additional matches to antigenic proteins. The sequence EDAK from Borrelia mapped to known antigen OspF (E-value: 4.6), and DYAFG from Syphilis maps to a lipoprotein in several strains of *Treponema pallidum* (E-value: 0.27). Malaria contained sequences (SNKQG, RLKEP) which both mapped to the ring-infect erythrocyte surface antigen (RESA) protein in *Plasmodium falciparum 3D7* (E-value: 0.072), and another sequence (DAFEY) mapping to one of the *pfEMP1* variants in *P. falciparum* (E-value: 3.5). The sequence FKEG

mapped to an MDR efflux protein in *Bordetella pertussis* (E-value: 3.5). These results are summarized in **Table 2.3**.

A

Sequence	Infection	Organism	Antigen	Known Antigen
AVHAD	Dengue	*Dengue virus (1 - 4)*	NS1	Yes (Y. Chen et al., 2010)
REGEK	Dengue	*Dengue virus 4*	Serine protease NS3	Yes (Garcia G, 1997)
DYAFG	Syphilis	*Treponema pallidium*	Lipoprotein	No
EDAK	Lyme's Disease	*Borrelia burgdorferii*	OspF	Yes (Pasternak & Dzikowski, 2009)
FKEG	Pertussis	*Bordetella pertussis*	Multidrug Resistance Protein	No
SNKQG/RLKEP	Malaria	*Plasmodium falciparum*	RESA-like protein	Yes (Anders, 1986; Gardner et al., 2002)
DAFEY	Malaria	*Plasmodium falciparum*	pfEMP1	Yes (B. Wagner et al., 2012)

B

Sequence	In IEDB	Membrane Protein	E Value	P Value
AVHAD	Yes	N/A	5.00E-04	0.0004
REGEK	Yes	N/A	8.30E-04	0.0007
DYAFG	No	Yes	0.027	0.026
EDAK	No	Yes	4.6	0.98
FKEG	No	Yes	3.5	0.96
SNKQG/RLKEP	No	Yes	7.20E-02	0.067
DAFEY	No	Yes	3.5	0.96

Table 2.3: Proposed Epitope Mappings for Disease Cohorts: Table of discovered epitope sequences and their proposed antigen mappings. The two dengue epitopes were previously verified using peptide tiling of the NS1 and NS3 proteins against Dengue sera. A further two (EDAK, DAFEY) map to well known and characterized antigens in *Borrelia burgdorferii* and *Plasmodium falciparum* respectively. The rest showed characteristics on the array consistent with epitopes, but map to proteins only discovered in high throughput experiments (hypothetical). E value refers to the expected number of matches to the presumed epitope sequence(s) within the proteome of interest, and the corresponding P Value refers to the chance of encountering at least one instance of the sequence within the proteome of interest. Not all proposed epitopes mapped to the proteome with significant P-values, but they are reported here as a "best guess" to explain the high response to these sequences on the arrays.

These sequences are short due to platform limitations, and the E-values for these matches varied based on the size of the proteome. The Dengue sequences are unlikely to arise by chance, with E-values of less than 10^{-3}. Likewise, the two matches to the RESA protein in *P. falciparum* together had a low E-value of 0.072 corresponding to a p-value of 0.067. These sequences were remarkably specific to a disease, and could possibly be used as a diagnostic **(Table 2.4)** or as leads for vaccines.

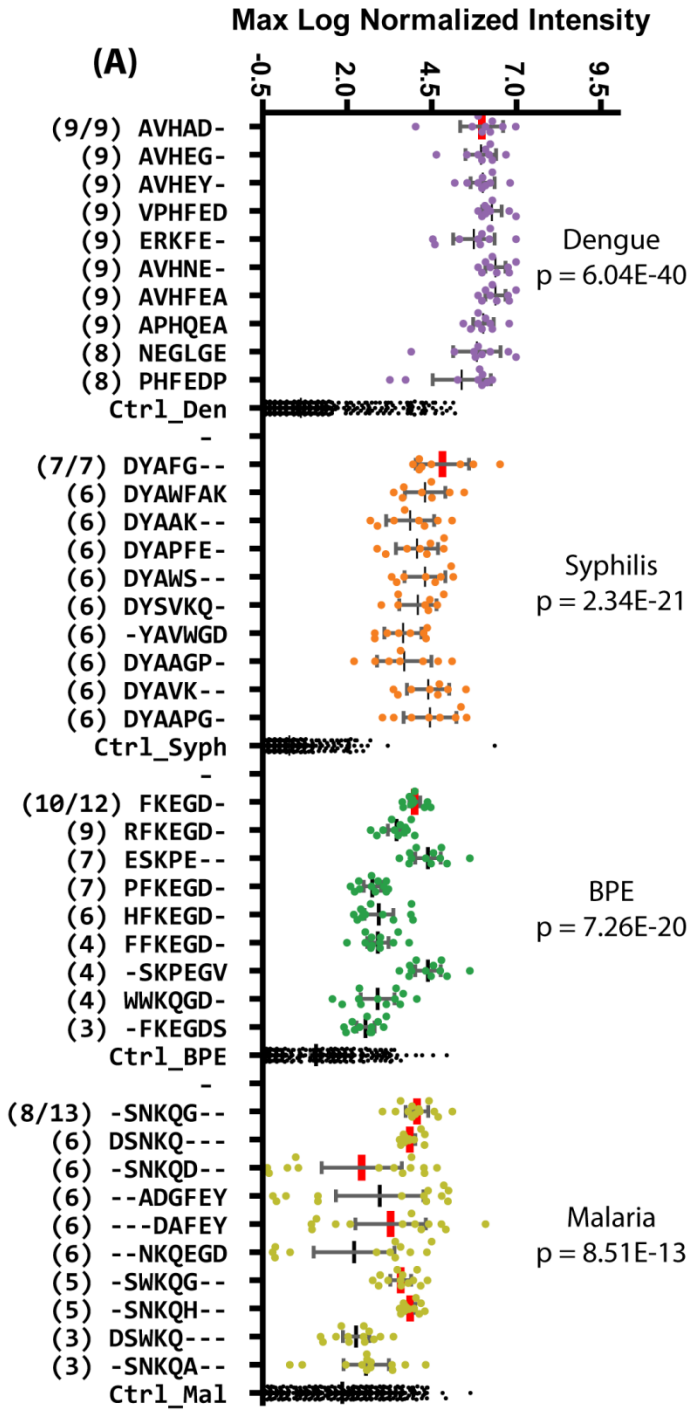
Sequence	Infection	Sensitivity	Specificity
AVHAD	Dengue	1	1
REGEK	Dengue	N/A	N/A
DYAFG	Syphilis	1	1
EDAK	Lyme's Disease	0.125	1
FKEG	Pertussis	0.83	1
SNKQG, RLKEP	Malaria	0.69	1
DAFEY	Malaria	0.46	1

Table 2.4: Sensitivity and Specificity of Epitope Candidates: Table of sensitivity and specificity calculations for the top epitope candidates from Table 2.3. The selection algorithm picks sequences such that sensitivity values are maximized, and may not be a reliable estimate of performance . Still, they do seem to map to antigenic proteins, and are remarkably specific to the cohort of interest. A blinded dataset is needed to achieve more reliable sensitivity and specificity values. Estimates for the REGEK sequence from Dengue could not be computed, as this was discovered using a separate set of arrays on which few samples were run.

Individual Epitope Determination in Patient Sera

In order to test the heterogeneity within disease groups, we asked which subsequences were differentially bound between an individual in a disease cohort versus normal. We found that epitope sequences revealed in the groupwise analysis were present

in most of the individuals from that group. All nine Dengue samples contained AVHAD as a significant subsequence. To visualize the extent of this overlap, we calculated the pairwise overlap of significant subsequences between individuals across disease groups (**Figure 2.4B**). This showed that individuals within the same group had a high degree of commonality with respect to the sequences seen by their immune system. These sequences were unlikely to appear in individuals from different disease groups, indicating these could be highly specific biomarkers for presence of a pathogen. The sequences with a high degree of overlap between individuals tended to have a very high fold change versus normal sera, while those with low overlap had lower fold changes.


Max Log Normalized Intensity
(A)
-0.5
2.0
4.5
7.0
9.5
(9/9) AVHAD-
(9) AVHEG-
(9) AVHEY-
(9) VPHFED
(9) ERKFE-
(9) AVHNE-
(9) AVHFEA
(9) APHQEA
(8) NEGLGE
(8) PHFEDP
Ctrl_Den
-
(7/7) DYAFG--
(6) DYAWFAK
(6) DYAAK--
(6) DYAPFE-
(6) DYAWS--
(6) DYSVKQ-
(6) -YAVWGD
(6) DYAAGP-
(6) DYAVK--
(6) DYAAPG-
Ctrl_Syph
-
(10/12) FKEGD-
(9) RFKEGD-
(7) ESKPE--
(7) PFKEGD-
(6) HFKEGD-
(4) FFKEGD-
(4) -SKPEGV
(4) WWKQGD-
(3) -FKEGDS
Ctrl_BPE
-
(8/13) -SNKQG--
(6) DSNKQ---
(6) -SNKQD--
(6) --ADGFEY
(6) ---DAFEY
(6) --NKQEGD
(5) -SWKQG--
(5) -SNKQH--
(3) DSWKQ---
(3) -SNKQA--
Ctrl_Mal
Dengue
p = 6.04E-40
Syphilis
p = 2.34E-21
BPE
p = 7.26E-20
Malaria
p = 8.51E-13

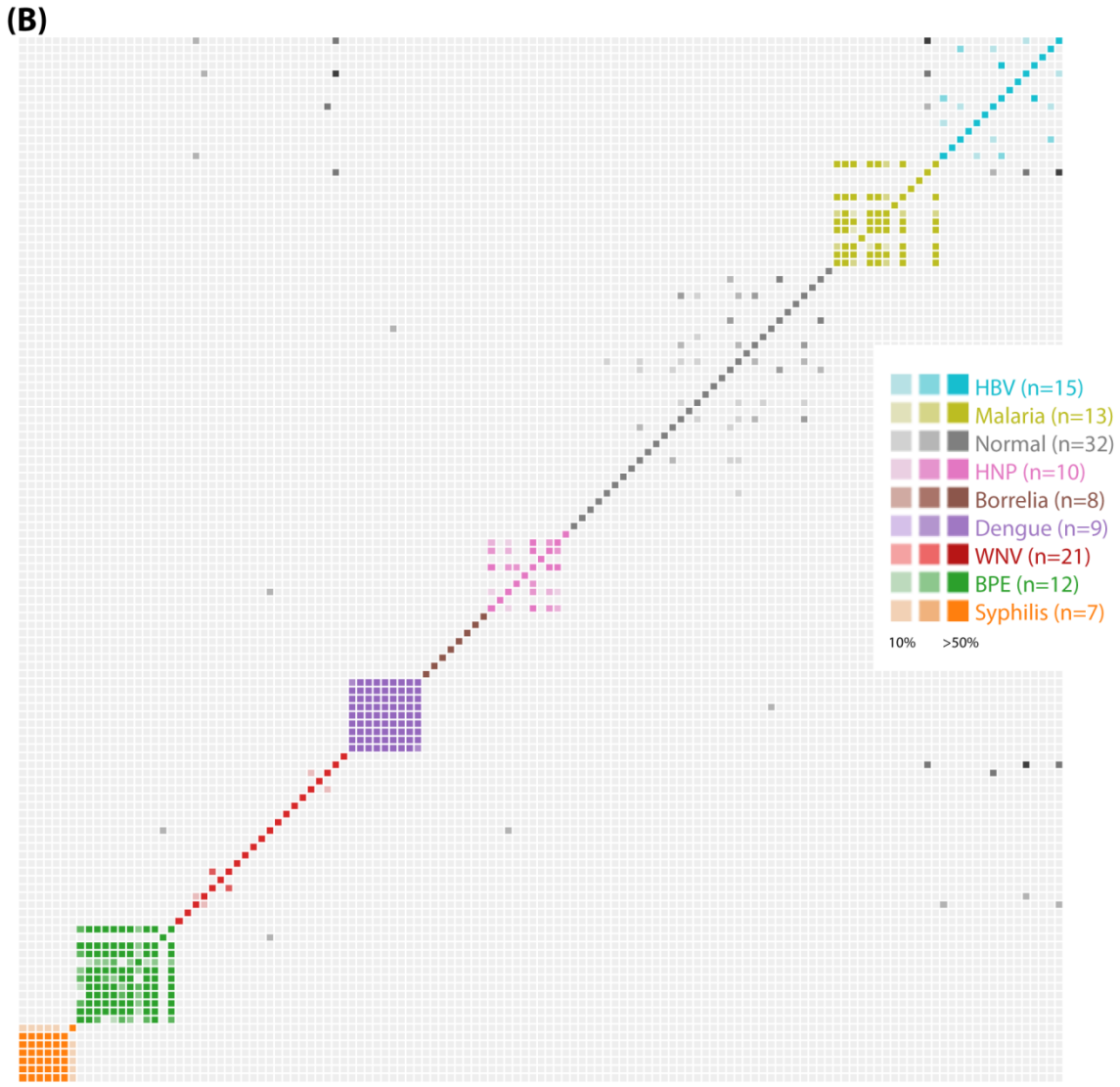


Figure 2.4: Top Significant Subsequences for Disease Cohorts: (A) Shows the top 10 most commonly appearing significant subsequences showing up in serum samples from the indicated disease cohorts. The number of patients within that cohort for which that sequence was called as significant is shown in parentheses on the left. The total number of samples in each cohort is given as a fraction next to the top significant subsequence and also on the figure legend to the right. Subsequences with exact matches to proteins within the pathogen are indicated with red bars (top sequences are listed in Table 2.3). Control data points show binding to the same collection of subsequences by the other tested cohorts. **(B)** Shows the pairwise fractional overlap in significant subsequences. A colored, saturated cell represents a pair of patients in the same cohort that shared at least 50% of their significant subsequences. Grayscale cells represent pairs of patients from different cohorts whose immune systems see similar sequences. Individuals within the same disease cohort show a much higher overlap between their significant subsequences than those in different cohorts or the normal cohort, indicating an association between the

discovered sequences and the disease state. BPE stands for Bordetella pertussis and HNP stands for Human Normal Pools, which consisted of pools of normal serum.

Additional Library Complexity Reveals Additional Epitopes

This assay relies on many simultaneous measurements of antibody/peptide interactions. It is useful to know how changes in library content affect results. Since only 27% of pentamers were represented on the original arrays, we hypothesized that a different random library would result in additional targets that were invisible to the original experiments due to not being present on the arrays. To test this, we created another array with a different set of 330,000 sequences. We then attempted to find epitopes using a dengue infected serum sample. This analysis revealed an additional epitope (REGEK, Dengue 4, E-value: $8.3x10^{-4}$) which was previously mapped in the IEDB and not present on the original array (**Figure 2.5**). This result shows that larger arrays could find additional antibodies that are already present in patient sera whose epitopes are not represented in smaller libraries. This argues for larger array libraries that would capture more sequences, revealing additional epitopes and giving a higher resolution picture of the immune response.

							FC	(n)
-	-	E	G	E	K	V	13.34	(3)
-	P	E	G	E	K		13.16	(10)
-	-	E	G	E	K	G	12.55	(36)
-	-	E	G	E	K	E	12.34	(8)
-	L	E	G	E	K		11.82	(6)
-	N	E	G	E	K	G	11.69	(3)
-	-	E	G	E	K	D	11.22	(8)
-	-	E	G	E	K		11.21	(57)
H	V	E	G	E	K		11.08	(3)
-	P	E	G	E	K	P	10.81	(3)
-	R	E	G	E	K		10.80	(13)
.	.	.	.	.	.	.		

							FC	(n)
E	D	W	K	Q			7.68	(3)
E	D	L	K	Q			6.38	(3)
E	D	N	K	E			6.12	(3)
E	D	A	H	G	F	K	5.69	(3)
E	D	S	K				5.48	(5)
E	D	N	K	R			5.42	(3)
E	D	A	K	Q			5.21	(3)
E	D	R	K	E			4.99	(3)
E	D	K	H				4.95	(3)
E	D	A	K	A			4.92	(5)
E	D	A	K				4.87	(22)
.	.	.	.	.	.	.		

Figure 2.5: Motifs found in Single Patients -- These motifs were associated with single patients within a disease cohort. (Left) was found in a Dengue patient and maps to NS3 (Garcia, 1997). It is a previously mapped epitope and observable on the peptide arrays. (Right) was present in a Borrelia patient and maps to the OspF protein, known to be associated with an immune response in dogs (Wagner et al., 2012). FC stands for fold change between the individual serum sample and a cohort of normal samples, and n refers to the number of peptides associated with that subsequences.

Mapping Epitope Information to a Database

We set out with the goal of being able to find epitopes in patient sera from unknown origin, and using that sequence information to identify the antibody eliciting proteins from a database. Having demonstrated that peptide microarrays are capable of resolving epitopes, we wanted to know if these sequences could predict the eliciting protein from a database of pathogen protein sequences.

Resolving a pathogen in a database given a few short sequences depends both on the size of the database and the length of the consensus motif. Using pairs of randomly generated sequences of varying lengths, we predict that a pair of pentamers if known exactly, or a pair of heptamers if known within 80% identity, are sufficient for resolving a pathogen in the Pathogen Proteome Database (**Figure 2.6**).

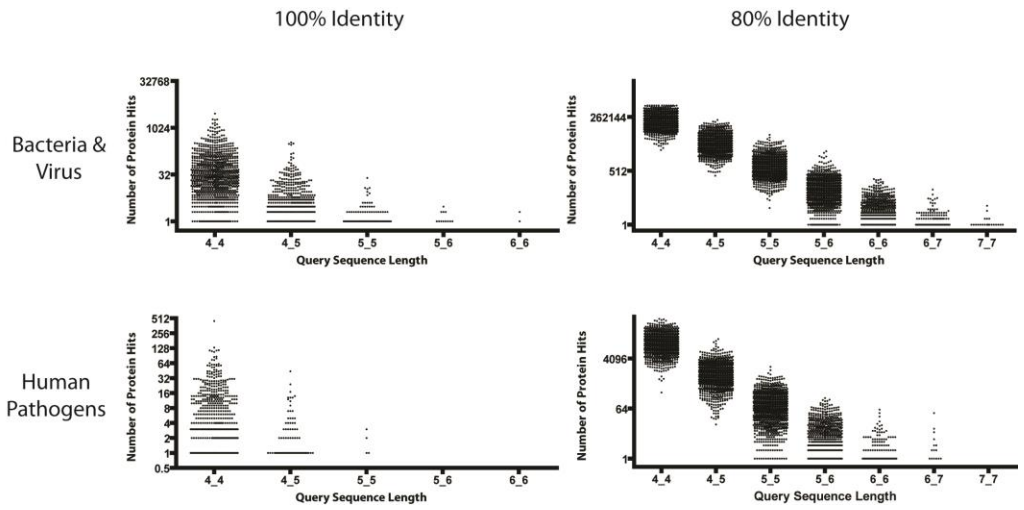


Figure 2.6: Finding Arbitrary Sequences in a Pathogen Database: These plots show the distribution of hits to pairs of arbitrary sequences of fixed lengths. Pairs of k-mers with specified lengths were drawn at random from the distribution associated with array sequences. These were searched against two databases: One containing over 4000 bacteria and viruses, and another containing 596 human pathogen strains. These plots show that when two 7-mer linear epitopes from the same protein antigen are known within 80% identity, unique pathogen identification should be possible.

Deciphering Eliciting Pathogen Proteins

To improve sensitivity, we opted for a restrictive search, relying on exact or near exact (80%) identity and matches in the same protein to multiple pentamer queries. Using significant subsequences from Malaria subjects, three epitope candidates were revealed (SNKQG, RLKEP, SNKQG). Searching these candidates against the Pathogen Proteome Database (multiple strains of each pathogen) resulted in uniquely identified membrane proteins from *P. falciparum* matching all three query sequences with 80% identity (**Figure 2.7**). Two of the query sequences matched with 100% identity to a RESA-like protein, a known antigen in *Plasmodium* infections. The probability of two randomly drawn pentamers matching to one or more proteins globally in this database of over 1 million sequences is<0.01.

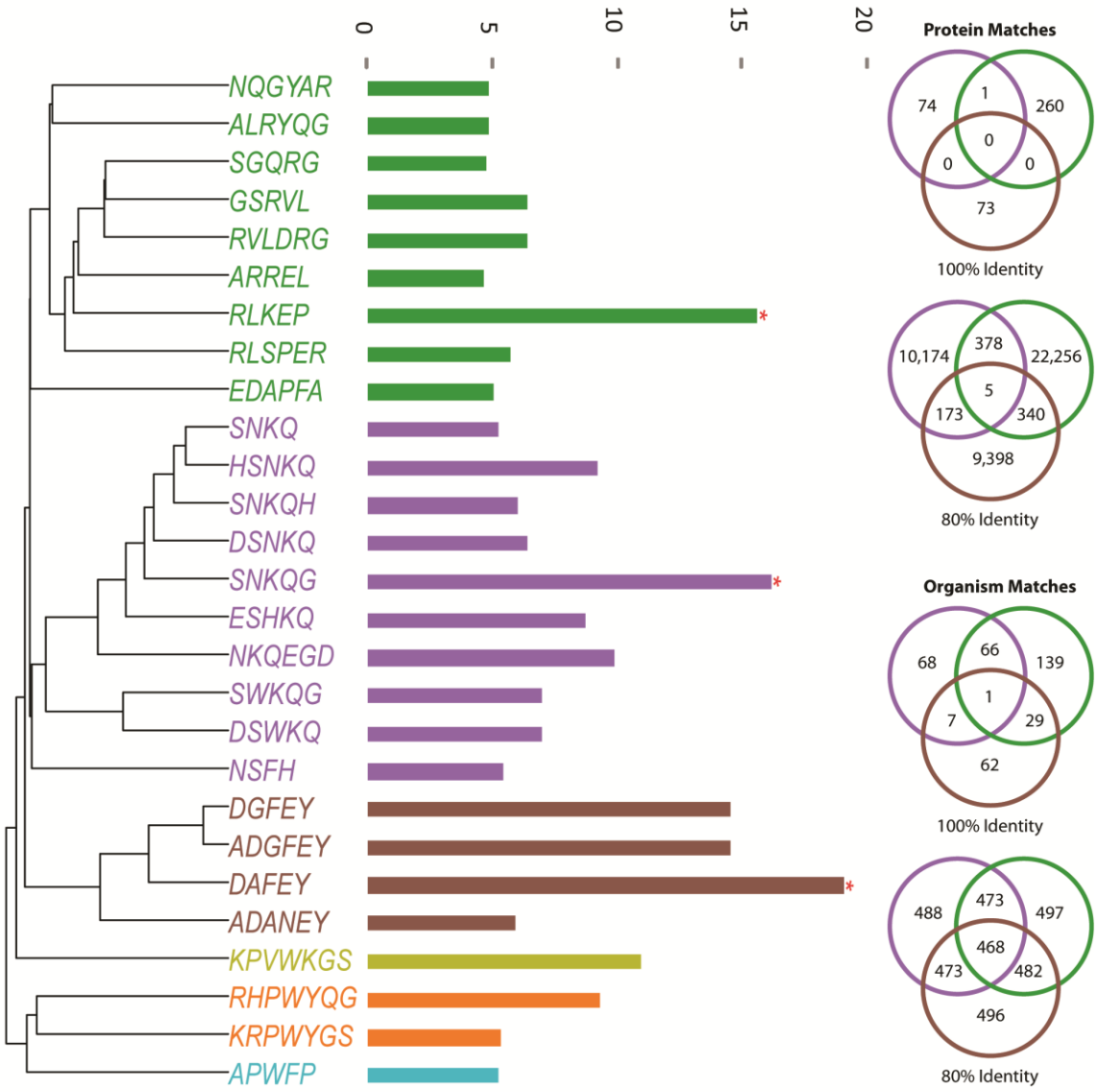


Figure 2.7: Using Significant Subsequences to Identify an Eliciting Pathogen: Sample specific significant subsequences from the Malaria cohort were combined, aligned and hierarchically clustered by single linkage. This revealed three distinct epitope candidates, indicated with red stars. These three sequences were queried against a database of 596 human pathogens for exact and 80% identity. Only one protein from P. falciparum out of all human pathogens contained both RLKEP and SNKQG. The probability of two array 5-mers hitting the same protein by chance is <0.001

Discussion:

First we asked whether random sequence arrays could resolve epitope sequences for well characterized monoclonal antibodies. We chose eight different monoclonal with well characterized reactivity to linear epitopes **(Table 2.1)**. The epitopes of 5 of 8 monoclonal antibodies were readily resolved. After verifying our method with monoclonals antibodies, we applied the technique to serum from cohorts infected with different human diseases. These samples were chosen to evaluate our ability to detect eptiopes across a broad range of pathogens. Two normal cohorts, one consisting of human normal samples and another consisting of pooled normal samples, were also included as controls. Epitopes consist with response to infection by Dengue, Malaria and possibly 3 other pathogens were identified. Finally, we compiled a list of human pathogen proteomes in order to test whether our discovered sequences were present in proteins from the pathogen of interest and whether pathogen identification in an uncharacterized serum sample is feasible.

The monoclonal experiments were designed to test whether 330,000 random-sequence peptides could correctly find a linear epitope. Peptide arrays are unique in that they provide binding information and also non-binding information, giving an overall picture of antibody specificity. Five monoclonals (HA, DM1A, 4C1, Ab1 and FLAG) bound only peptides that were related to their target. P53Ab1 essentially bound a single sequence RHSVV, did not tolerate substitutions, and did not cross react with additional peptides. HA, 4C1, FLAG, and DM1A tended to allow substitutions in certain positions to varying degrees dependent on the sample. P53Ab8 bound sequences similar to the epitope, but these were overshadowed by sequence dissimilar distracter sequences

(**Figure 2.3A)**. Two antibodies (A10, 2C11) seemed to only bind sequence dissimilar peptides. These differences in apparent binding may reflect true variation in antibody cross reactivity characteristics, or could be a side-effect of choosing peptides randomly. Further studies with additional antibodies are needed to determine the extent to which the arrays can predict antibody specificity. Given the importance of monoclonal antibodies in the therapeutic pipeline (Leavy, 2010; Reichert & Valge-Archer, 2007), a quick way to screen out undesirable cross reactions on a simple, high throughput platform is desirable.

In agreement with previous studies using dense peptide arrays (Edfors et al., 2014; Forsström et al., 2014), the antibodies bound a variety of sequences, many of which had little or no relationship with the true epitope. This was the impetus for the subsequence approach, which was successful in filtering out these non-specific sequences in the five monoclonal antibodies (**Figure 2.1**) and made obvious the most significant binding motifs.

These motifs, despite being pentamer sequences with only three to five amino acids in common with the eliciting peptide, bound very strongly to their targets, often over twenty fold above background (**Figure 2.2)**. This strong, specific binding suggests that epitopes only require a limited number of binding residues, a phenomenon also observed in previous studies (Buus et al., 2012a; Sivalingam & Shepherd, 2012). There is likely an evolutionary optimization between simplicity (low number of binding residues) and specificity (the need to recognize a unique target). Given that the size of sequence space increases exponentially with the number of residues, an antibody does not require many residues to maintain global specificity to a target.

Previously, we attempted epitope mapping on smaller arrays with 10,000 peptides, with modest success for monoclonal antibodies but no predictive power in the case of patient sera (Halperin et al., 2011). These data show that this was most likely due to a sparse representation of peptide sequences, with only 0.5% of pentamers represented in triplicate. The arrays used in this study provided a much denser sampling of this space, with 27% of pentamers represented. This improved sampling corresponded to an improvement in resolving epitopes in patient sera.

Dengue samples in particular seemed to react strongly to a particular epitope on the NS1 protein, shared by many strains of the virus. Because this is shared between strains, this antibody is likely non-protective, and serves to distract the immune system. This explains why this was seen in all patients tested when they were likely infected with different strains of the virus.

Malaria was an interesting cohort because multiple sequences (SNKQG, RLKEP) mapped to the Ring Erythrocyte Surface Antigen (RESA) protein in *P. falciparum*. This protein is associated with the membrane of newly invaded erythrocytes (Brown et al., 1985; Foley, Tilley, Sawyer, & Anders, 1991), is an important virulence factor that facilitates erythrocyte attachment to blood vessel epithelium, and presents a tempting target for the immune system. The *P. falciparum* proteome is so large that it would be almost impossible to map the eliciting protein from a single pentamer, but in this case two peptides mapped to the RESA protein, improving the likelihood of a true match. A further sequence (DAFEY) was found in six samples and mapped to a PfEMP1 protein, which is one of a family of variant antigens associated with infected erythrocytes and thought to be important mechanisms for immune system distraction and evasion (Su et

al., 1995). Expression of these proteins is constantly being switched in order to evade the host immune response, and it is likely more antibodies against this family would be found in a larger study.

The Syphilis and Bordetella cohorts also showed consensus sequences which mapped to proteins, but the annotations on these are less comprehensive, and it is unknown if they are antigenic. They do appear to be surface associated proteins, but they are hypothetical and direct studies about their expression or function have not been reported in the literature.

While many individuals within a cohort shared (possibly non-protective) epitopes, heterogeneous responses were also observed. Two Borrelia samples bound the consensus sequence EDAK. While this is too short to call conclusively, it maps precisely to the OspF protein found in several strains of the bacterium. This is a proven antigen(B. Wagner et al., 2012), and the subsequence is found in a region between two trans-membrane sections of the protein, thus it is a feasible location for an epitope. In some cases while the assignment may not be definitive, it may allow a reduction to likely candidates.

The presence of homogenous epitopes within cohorts is promising, as these arrays were originally developed to monitor serum and predict the presence of the disease as part of a diagnostic platform. Previously we had shown that this assay is capable of capturing a "signature" of the immune system, aiding in the diagnosis of disease (Hughes et al., 2012; Kukreja et al., 2012; Legutki et al., 2010; Restrepo et al., 2012; Restrepo et al., 2011b; Sykes et al., 2013). While machine learning algorithms can accurately classify serum samples into the correct disease category, until now we have not shown that these

signatures contain epitope sequences. The serum samples revealed patterns consistent with those seen in the predictive monoclonal samples, and appear to map to antigenic proteins from the pathogen **(Table 2.3)**. In the case of the two dengue epitopes, validation that these sequences are indeed antibody targets has previously been done by other groups (Y. Chen et al., 2010; Garcia, 1997), but this has not been completed for the other sequences, and for now they should be considered putative candidates.

As previously mentioned, the arrays contain around 27% of possible pentamers in triplicate. Given this modest representation, one would predict approximately a one in four success rate when mapping epitopes. However, in both the monoclonal and serum samples, success rates were much higher with discernable epitopes revealed in over half of tested samples/cohorts. One likely explanation is that infected sera contain multiple antibodies each with unique specificities, and only a subset was "visible". We saw some evidence of this when we repeated the assay in Dengue on a new array, which revealed an additional validated epitope in previously unrepresented space.

Identifying eliciting proteins using sequence information gleaned from the arrays with the current 330K peptides per array is challenging. These arrays contain a relatively limited amount of sequence information compared to what is available in genome or transcriptome annotation studies. A typical BLAST search of a pentamer against a database of human pathogens is likely to be dominated by spurious and insignificant results. The arrays tend to reveal only consensus motifs that are present on the array, and not exact sequences. Second, the array only provides ample coverage of sequence space up to five amino acids, limiting the lengths of epitopes that can be reasonably discovered. Even given these limitations we have demonstrated that it is possible to identify likely

antigenic proteins using combinatorial random sequence peptide arrays. Interestingly, epitope candidate pentamers gleaned from the arrays were much more likely to match pathogenic protein sequences than randomly drawn array pentamers (data not shown). This indicates that epitopes are actually much less diverse than random or even life-space sequences, supporting the idea that antigen space is intrinsically convergent (Campo et al., 2012). It is also fairly straightforward to increase the number of peptides many fold.

The techniques underlying this technology are highly amenable to high-throughput manufacturing. Given that we identified different epitopes by using two different libraries, is likely larger arrays would achieve the sensitivity required for *apriori* pathogen identification. The approach seems promising in that true epitopes were revealed along with several previously undiscovered linear sequence segments in antigenic proteins. Such an approach could help identify antigenic hot spots within proteins and immunodominant epitopes with high resolution using an assay that is significantly cheaper in time and labor than display techniques, facilitating high throughput screening of serum and monoclonal antibodies.

CHAPTER 3

IMMUNOSIGNATURES FOR DENGUE DIAGNOSTICS

This chapter contains significant contributions from Xiao Wang, who designed the original study and collected most of the array data (CIM10K). My contribution was to design the classification methodology using best practice machine learning principals and make it available for others to use. I evaluated Xiao's raw data using this methodology and wrote the results including figures. I also included additional data from new, larger arrays (CIM330K), using this to identify Dengue epitopes and evaluate the WHO subtype samples.

Abstract

Dengue Fever and Dengue Hemorrhagic Fever represent emerging infections affecting many nations throughout the world. Though accurate diagnostics are essential for epidemiological studies and tracking, existing FDA approved methods are hampered by dengue antigen cross reactivity in sera from individuals with other infections. In order to overcome this problem, we tested the ability of an antibody profiling technology based on non-natural sequence peptide microarrays called immunosignatures to distinguish primary and secondary Dengue infection from infections of Malaria , West Nile Virus and from non-infected people using sera. Classification using support vector machines and a rigorous cross-validation scheme resulted in good metrics for distinguishing between Dengue and other diseases (AUROC: ~0.92, sensitivity: ~0.96, specificity: ~0.76), and also perfect classification of a small number of primary versus secondary infections. Further analysis revealed cohort specific amino acid biases and consensus

sequences to known and unknown Dengue epitopes. Immunosignature microarrays may be a promising technique for Dengue diagnostics and epitope mapping.

Introduction

Dengue fever (DF) and Dengue hemorrhagic fever (DHF) are emerging infections in over one hundred tropical and subtropical nations around the world. It is estimated that there are ~400M infections per year (Murray, Quam, & Wilder-Smith, 2013). Accurate and prompt diagnostics are important for epidemiological tracking and evaluation of Dengue epidemics. Since secondary infection is a risk factor for DHF (Kliks, Nisalak, Brandt, Wahl, & Burke, 1989; Rigau-Pérez et al., 1998), diagnostics that can separate primary from secondary infection are of principal importance. Over the past several years, several IgM and IgG ELISA-based assays have been developed with useful sensitivity and specificity for both predicting infection and differentiating primary versus secondary infection (Guzmán & Kourí, 2004). Despite these promising advances, these ELISA-based diagnostics for Dengue often show cross reactivity to other flaviviruses, such as West Nile Virus (WNV), Japanese Encephalitis Virus (JEV) and Yellow Fever Virus (YFV) (Banoo et al., 2008; Hunsperger et al., 2009; Schwartz, Mileguir, Grossman, & Mendelson, 2000), and possibly even Malaria antigens (Hunsperger et al., 2009). In order to overcome these deficiencies, we tested a peptide microarray consisting of over 10,000 peptides chosen from random sequence space as a platform for specifically differentiating dengue from other related disease and uninfected subjects.

Peptide microarrays have proved useful for measuring changes in the humoral immune system. These multiplexed assays, known as immunosignatures (Sykes et al., 2013), measure antibody binding to thousands of peptide sequences in a single assay, and

have been employed for predicting vaccine efficacy (Legutki et al., 2010), distinguishing between related pancreatic diseases (Kukreja et al., 2012), mapping monoclonal epitopes (Halperin et al., 2011), and mapping epitopes from infectious disease sera (manuscript submitted). Unlike most assays which measure a single biomarker target, immunosignatures seek to capture the unbiased signature of the humoral immune system and to associate this pattern with a disease using well understood machine learning algorithms. It is a single method for multiple diagnostics, and as such may not be as susceptible to cross reactivity as ELISA-based assays or assays using natural peptide sequences (manuscript submitted).

Methods

Serum Sources

Samples tested here include Dengue (Seracare PVD-201), Malaria (Seracare DS-774) and West Nile Virus (Seracare, various) as well as Non-Infected samples collected from laboratory volunteers and randomly chosen for testing. See **Table 3.1** for details.

Disease	Sample	Min	Max	Range	Mean	Country of Collection
Dengue	201-01	220	65535	65315	1540.0	Colombia
Dengue	201-02	202	65535	65333	2383.1	Honduras
Dengue	201-03	254.5	53863.5	53609	2572.1	Honduras
Dengue	201-04	175.5	56359	56183.5	1445.8	Honduras
Dengue	201-05	201	56124	55923	1469.3	Honduras
Dengue	201-06	212	65535	65323	1589.2	Honduras
Dengue	201-07	330	62266.5	61936.5	2084.5	Colombia
Dengue	201-08	275	65535	65260	3367.4	Honduras
Dengue	201-09	239	65535	65296	2242.8	Honduras
Dengue	201-10	218	65535	65317	3181.6	Ecuador
Dengue	201-11	225	65535	65310	1604.8	Honduras
Dengue	201-12	213.5	65535	65321.5	1435.3	Honduras

Dengue	201-13	180.5	65535	65354.5	1147.4	Honduras
Dengue	201-14	227.5	61502	61274.5	1432.9	United States
Dengue	201-15	194	65254.5	65060.5	1428.5	Honduras
Dengue	201-16	190	65535	65345	2408.6	Ecuador
Dengue	201-17	191	62409.5	62218.5	1196.8	Colombia
Dengue	201-18	257.5	65535	65277.5	1856.5	Honduras
Dengue	201-19	212	64339.5	64127.5	1913.0	Ecuador
Dengue	201-20	183.5	65535	65351.5	3085.2	Ecuador
Dengue	201-21	214.5	65535	65320.5	1527.9	Ecuador
Malaria	HP-10	160.5	63018.5	62858	4408.8	Honduras
Malaria	HP-11	440.5	63008.5	62568	10119.7	Honduras
Malaria	HP-17	93.5	14290	14196.5	775.6	Honduras
Malaria	HP-7	155	49734	49579	3490.1	Honduras
Malaria	HP-8	162.5	41997	41834.5	2876.2	Honduras
Malaria	HP-9	165.5	42974.5	42809	3300.7	Honduras
Malaria	MA-34	257.5	65535	65277.5	2492.0	Honduras
ND	ND-134	204	65535	65331	1478.5	US
ND	ND-149	283.5	65535	65251.5	2648.6	US
ND	ND-153	211.5	61388	61176.5	2060.6	US
ND	ND-154	209.5	65535	65325.5	1533.6	US
ND	ND-155	251.5	65535	65283.5	1939.2	US
ND	ND-157	185	48285.5	48100.5	1286.0	US
ND	ND-158	322.5	60912.5	60590	2006.4	US
ND	ND-159	289	60124	59835	1896.5	US
WNV	WNV-2	203	41665	41462	1956.3	Unknown
WNV	WNV-3	190	65535	65345	1589.8	Unknown
WNV	WNV-1	275	65535	65260	2115.7	Unknown
WNV	WNV-4	200	45378	45178	1736.1	Unknown
WNV	WNV-9147710	299	62588	62289	2225.8	Unknown
WNV	BMI-140433	287	65535	65248	3458.0	Unknown
WNV	BMI-141170	222	65535	65313	1623.2	Unknown
WNV	BMI-145662	235.5	65535	65299.5	2776.6	Unknown

Table 3.1: Descriptive Statistics for Each Sample: Twenty one Dengue samples, seven Malaria samples, eight Non-Infected samples, and eight West Nile samples were run on the array. Dengue, Malaria and West Nile samples were collected from Colombia, Ecuador, United States and Honduras while Non-Infected samples are non-endemic and collected from volunteers at Arizona State University. WNV stands for West Nile Virus

and ND stands for Non-Infected Donor. The metrics show the minimum, maximum and dynamic range of measurements collected for the IgG based assay.

Array Sources

CIM10K arrays consists of 10,000 peptides (Sigma-Aldrich, St. Louis, MO) printed on a glass slide. These arrays contain a fixed number of peptides, generated to cover as much as possible the space of linear peptide sequences, thus maximizing the chances of representing an arbitrary linear epitope. For epitope determination studies a new array produced by Healthtell, Inc (Chandler, AZ) (HT330K) containing over 330,000 features was used. HT330 peptides average 12 amino acids in length.

Primary v Secondary Infection Determination

Dengue samples were determined either as primary or secondary infections based on the results of the Panbio ELISA IgG/IgM Duo Assay and were all from Ecuador where the virus has been endemic since the 1980s (San Martín et al., 2010).

Assay Conditions

The assay conditions for the CIM10K arrays have been described in (Legutki et al., 2010) and for the HT330K arrays in (Legutki et al., 2014). Briefly, the process for the CIM10k is:

a. Slides are first washed in a pre-wash solution (7.73% acetonitrile, 33% isopropanol, 0.55% TFA), then thrice washed in 1xTBST followed by three washes in ddH20, then dried by centrifuge.
b. Incubation is handled using the Tecan Automated Slide Processing System. Slides are blocked for 1 hr in blocking solution (30% BSA, 6.9 uL mercaptohexanol, 25uL Tween 20 in 50mL PBS). Slides are then treated with 1:500 sera diluted in

blocking buffer (30% BSA, 25uL Tween 20, 50mL PBS) for 1 hr, followed by 5nM goat anti-human IgG in blocking buffer for 1hr, followed by 5nM Alexaflour647 conjugated streptavidin for 1hr.

c. Slides are scanned using an Innoscan 910 Microarray Scanner (Innopsys; Carbonne, France) at appropriate wavelengths matching emission of the tertiary dye.

For the HT330 array the process is:

a. Prior to assay, arrays are washed in 100% DMF for one hour, then introduced to PBST incubation buffer (3% BSA in Phosphate Buffered Saline, 0.05% Tween 20) over a period of six hours to allow the solvent phase to be completely transitioned to aqueous phase. The arrays are then processed by incubating in the presence of antibodies or serum and detected by fluorescent antibody.

b. Residual DMF is removed by two 5 min washes in distilled water. Arrays are equilibrated in PBS for 30 min and blocked in incubation buffer (3% BSA in Phosphate Buffered Saline, 0.05% Tween 20 (PBST). Arrays are washed and briefly spun dry prior to loading into the multi-well gasket (Array-It, Santa Clara, CA). Incubation buffer is added to each well (100ul) and 100 ul of 1:2500 diluted sera is added for a final concentration of 1:5000. Arrays are incubated for 1hr x RT with rocking then washed with PBST and 1% BSA in PBST using a BioTek 405TS plate washer.

c. Anti-human IgG-DyLight 549 (KPL, Gaithersburg, MD) is added to a final concentration of 5.0 nM. Following 1hr x RT with rocking, unbound secondary is removed by washing in PBST followed by washing in distilled water. The arrays

were removed from the gasket while submerged, dunked in isopropanol and centrifuged dry (800xG, 5 min). Arrays are scanned at 533nm using an Innoscan 910 array scanner (Innopsys, Carbonne, France). Features were aligned and extracted using GenePix Pro 6.0 (Molecular Devices, Sunnyvale, CA).

Alignment, Data Extraction and Normalization

Slides were aligned and data extracted using GenePix Pro 6.0, and slides with >80% correlation between replicates were chosen for analysis. Slides with poor correlation were assayed again. Per-slide fluorescence intensities were normalized to the 50^{th} percentile in order to remove variation caused by small differences in dye concentration or PMT and laser variation.

Cross Validation

All classification results reported here were "leave one from each class out" cross validated. One instance for each class was randomly selected and set aside for model testing. The rest were used for feature selection and model training. This process was repeated iteratively until the class membership of each instance has been predicted at least once.

In order to rule out over-fitting as a driver of classification results, this same procedure was repeated with randomly shuffled class labels. This means that for each column in the array matrix, the true class label was removed and replaced with a randomly chosen class label. Proportions for each class were kept the same as those from the true labels.

Feature Selection

Selecting peptides relevant to a particular disease of interest is of principal importance for immunosignatures. T Tests between relevant groups (Dengue v Non-Infected, Primary v Secondary, Dengue v Non-Infected, Malaria, WNV) are used to select peptides that are significantly different between groups (p values typically were 10^{-6} or less for significant peptides). P values were Benjamini and Hochberg FDR corrected to account for multiple hypotheses. All significant features were used for classification.

SVM Classification

The SKLearn package (Python 2.7) was used to train and test linear kernel SVM models. Randomized cross validation was employed where the dataset is repeatedly split into training and test sets at random such that the test set contains one sample from each cohort (leave one from each class out). Then, peptides are selected and an SVM model is trained using only data from the training set. The model then predicts the test set, producing sensitivity and specificity estimates. To produce a negative control on this procedure, the same steps were repeated on a dataset with the class labels (DENGUE, NON-INFECTED, etc) randomly shuffled. The procedure for both the true and shuffled labels was repeated 100 times, and the mean specificity and sensitivity values were used as the classification result. The program used for classification can be downloaded at https://github.com/joshuaar/CIM-Scripts/blob/master/classif2.py. Confidence intervals were estimated using a bootstrapping procedure with the mean calculated from 10,000 bootstrap samples

from the cross validation results. The 5th and 95th quantiles from this calculation samples are reported as the confidence interval.

Sequence Enrichment for Epitope Determination

Enriched subsequences from the signatures within the highest binding peptides were found and used to identify conserved motifs. The top 500 peptides by normalized fold change between a dengue sample and the non-infected cohort were searched for enriched subsequences (k-mers) of length five or greater. Enrichment probabilities were approximated with the binomial distribution. All enrichments with $P < 0.005$ were selected and used to produce motif cartoons with WebLogo(Crooks, Hon, Chandonia, & Brenner, 2004). Code for calculating enrichments can be downloaded at https://github.com/joshuaar/CIM-Scripts/blob/master/CalcEnrichments.py.

Results

Dengue infected versus non-infected sera (IgG)

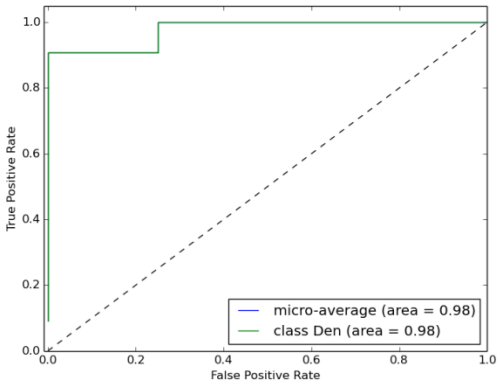
The first and most basic question was whether IgG based immunosignatures can distinguish Dengue positive sera from non-infected sera.. Twenty one Dengue serum samples collected from Colombia, Honduras, Ecuador and the United States were compared to eight non-infected samples collected from the United States. The classification procedure was applied to these data, and **Figure 3.1** shows the cross-validated results for this test. Overall accuracy of ~0.90 (CI: 0.87, 0.94) was achieved using this method, with approximately 0.83 specificity (CI: 0.77, 0.89) and 0.98 sensitivity (CI: 0.93, 0.99). The same procedure applied to the same data with randomly shuffled labels resulted in classification no better than random, with an overall accuracy of ~57% in balanced test sets, indicating our method is not over fitting the data.

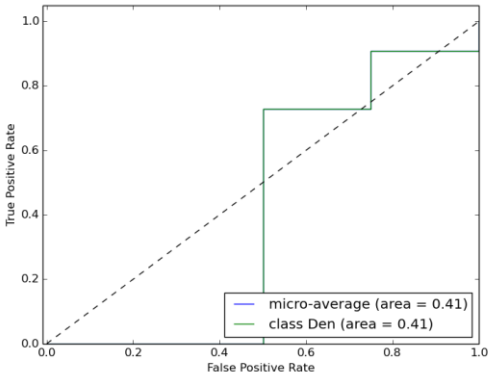
Dengue vs. non-infected sera (IgM)

IgG is a specific biomarker for Dengue, but often titers are low in primary infections and IgG is seldom used in existing diagnostic techniques. We tested an IgM based immunosignature for classification accuracy between dengue and non-infected sera. Accuracy was slightly higher in this test, averaging 0.95 (0.94 specificity, 0.96 sensitivity). Shuffled labels yielded an accuracy of ~55%, indicating over-fitting is unlikely. These results are summarized in **Figure 3.1**. It is interesting to note that the peptides used for IgM classification are different from those used in IgG classification (P=0.79).

3.1A

True Labels **Randomly Shuffled Labels**





Dengue vs. Non-Infected (IgG)	Specificity	Sensitivity	Accuracy
True Labels	0.83 (CI: 0.77, 0.89)	0.98 (CI: 0.95, 1.0)	0.905 (CI: 0.87, 0.94)
Randomly Shuffled Labels	0.13 (CI: 0.09, 0.17)	0.79 (CI: 0.75, 0.82)	0.57 (CI: 0.54, 0.59)

Dengue vs. Non-Infected (IgM)	Specificity	Sensitivity	Accuracy
True Labels	0.94 (CI: 0.90, 0.98)	0.96 (CI: 0.93, 0.99)	0.95 (CI: 0.93, 0.98)
Randomly Shuffled Labels	0.15 (CI: 0.09, 0.21)	0.96 (CI: 0.92, 0.99)	0.55 (CI: 0.53, 0.59)

3.1B

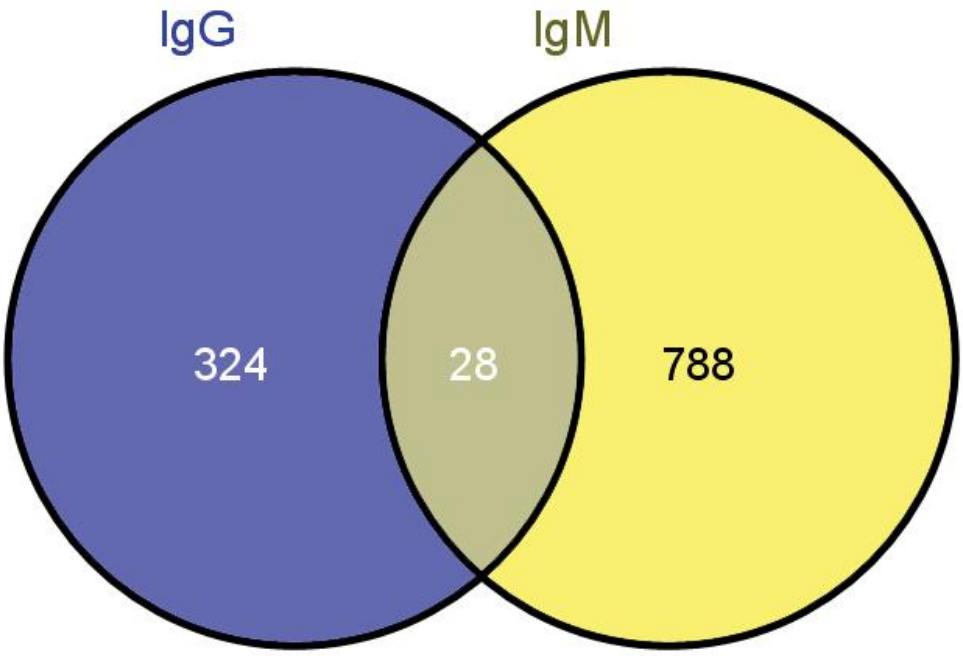


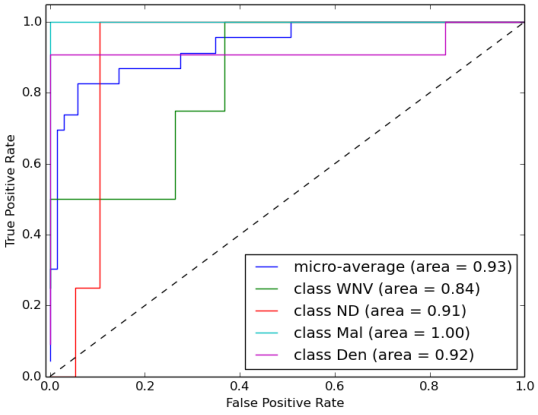
Figure 3.1: Dengue vs. Non-infected Classification Results – (3.1A) Top shows ROC curves for the binary classification scheme (Dengue vs. non-infected). Average and standard deviations of sensitivity, specificity and accuracy are reported from leave one from each class out cross validation. Shuffled labels refer to the fact Dengue and Non-Infected labels were shuffled randomly such that the data no longer represent true class memberships. This is a way to rule out model over fitting as a driver of classification results. **(3.1B)** Overlap between peptides selected for the IgG based assay and those selected for the IgM based assay (FDR Corrected $P < 0.05$). Under the hypergeometric distribution, the overlap is not significant ($P=0.79$) indicating that IgM and IgG assays use distinct sets of peptides.

Dengue Versus Non-Infected, WNV, Malaria (IgG)

Encouraged by the single class performance of the assay, additional cohorts of West Nile Virus and Malaria infected sera were added and classified using the same procedure. Overall accuracy fell from 0.90 to 0.79 (CI: 0.75, 0.82), but was still well above the uninformative accuracy of 0.25. For Dengue, sensitivity in this test was 0.96 (CI: 0.93, 0.99) and specificity was 0.76 (CI: 0.73, 0.80). Shuffled labels only resulted in an accuracy of 0.23, which is in line with the expected value for random data.

ROC curves and sensitivity/specificity estimates for each class are shown in **Figure 3.2**. These are substantially better than what could be expected by random chance and as in the binary classification result, the cross validation procedure appears rigorous as it cannot classify randomized labels better than random chance. Details on selected peptides are given in **Table 3.2.**

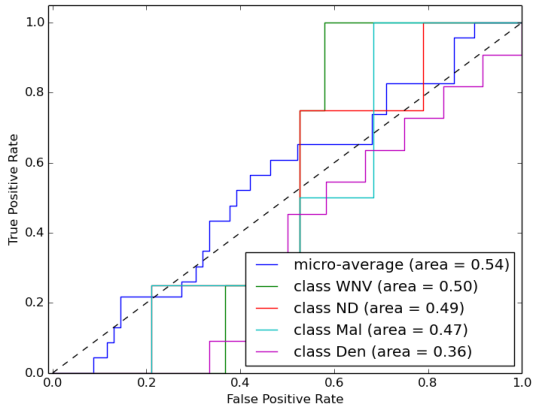
True Labels



Disease	Specificity	Sensitivity
WNV	0.99 (CI: 0.98, 0.99)	0.47 (CI: 0.39, 0.55)
ND	0.97 (CI: 0.95, 0.98)	0.8 (CI: 0.73, 0.86)
Mal	0.98 (CI: 0.97, 0.99)	0.91 (CI: 0.86, 0.95)
Den	0.76 (CI: 0.73, 0.80)	0.96 (CI: 0.93, 0.99)

Overall Accuracy: 0.79 (CI: 0.75, 0.82)

Randomly Shuffled Labels



Disease	Specificity	Sensitivity
WNV	0.92 (CI: 0.89, 0.95)	0.03 (CI: 0.01, 0.06)
ND	0.90 (CI: 0.88, 0.93)	0.08 (CI: 0.04, 0.13)
Mal	0.92 (CI: 0.89, 0.95)	0.08 (CI: 0.04, 0.13)
Den	0.23 (CI: 0.19, 0.27)	0.76 (CI: 0.69, 0.83)

Overall Accuracy: 0.23 (CI: 0.21, 0.26)

Figure 3.2: Multidisease SVM ROC curve for real labels and shuffled labels: **(3.2A)** A multiclass SVM was trained and tested with four classes: West Nile Virus (WNV), non-infected Non-Infected Donor (ND), Malaria (Mal), and Dengue (Den). The left side ROC indicates real labels and predicts performance of the model on new samples. The right side is the SVM tested on randomly shuffled class labels, and should be unpredictive. These curves show there is a real difference between random labels and true labels, indicating the model is not over-fitted.

Significant Peptides for Each Class P<0.05

Disease	Uncorrected	FDR	Bonferroni
Dengue	4269 ± 400	1024 ± 723	134 ± 150
Malaria	4869 ± 140	1253 ± 181	237 ± 56
WNV	1279 ± 160	6 ± 14	5 ± 7
Non-Infected	1384 ± 280	1 ± 1	3 ± 2

Table 3.2: Number of Peptides Selected Using Different Multiple Testing Correction Procedures: Data was split into training and test sets such that one instance from each class was placed in the test set. Features were selected from the remaining instances using Welch's T-Test. The number of peptides significant by this test varied depending on which multiple testing correction was used. Uncorrected refers $P < 0.05$, FDR refers to Benjamini and Hochberg FDR < 0.05, and Bonferroni refers to $P < 5.26 \times 10^{-6}$. Dengue and Malaria produced the most significant peptides, and were also the best classified (AUC Dengue = ~0.92, AUC Malaria = ~1.0).

ELISA Results (IgG, IgM)

Panbio ELISA results were collected from Seracare and were used to call primary versus secondary infections **(Table 3.3)**. Three samples showed a low IgG response and a high IgM response, characteristic of a primary infection. These predicted primary samples were used in a further classification scheme for diagnosing primary versus secondary infection.

Significant Peptides (P<0.05) For Dengue Subtypes

Sample	Uncorrected	FDR	Bonferroni
Dengue 1	7065	5996	8
Dengue 2	574	6	3
Dengue 3	3957	0	0
Dengue 4	755	8	1
Non-Infected	1147	5	3

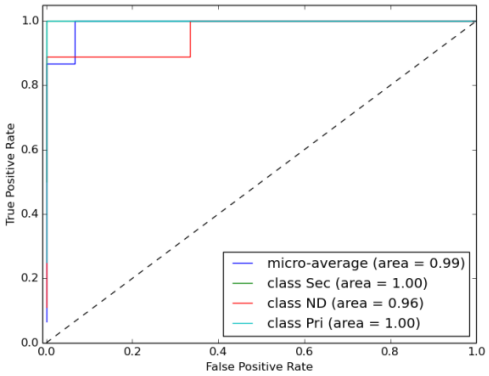
Table 3.3: Significant Peptides Between Dengue Subtypes: Subtype prediction is an important problem in Dengue diagnostics. In order to test whether peptides could differentiate subtypes, peptides were selected with Welch's T Test and multiple testing corrections were applied. Uncorrected refers $P < 0.05$, FDR refers to Benjamini and

Hochberg FDR < 0.05, and Bonferroni refers to $P < 5.26 \times 10^{-6}$. All subtypes showed significant peptides in the strict Bonferroni multiple testing regime except for Dengue 3.

Dengue Primary versus Secondary Infection Prediction (IgG)

The three predicted primary samples were labeled appropriately and the SVM model was trained on these newly labeled classes. ROC curves along with sensitivity and sensitivity estimates for primary versus secondary are shown in **Figure 3.3**. These gave almost perfect AUROC scores, achieving 1.0 for secondary vs. primary and non-infected, 0.94 for non-infected vs. primary and secondary, and 0.98 for primary vs. non-infected and secondary. Sensitivity for detecting DHF high risk secondary infection from non-infected and primary was 0.8 (CI: 0.72, 0.85) and specificity was 0.89 (CI: 0.86, 0.93). Again the cross validation scheme appears valid as the procedure with randomly shuffled labels non-predictive results.

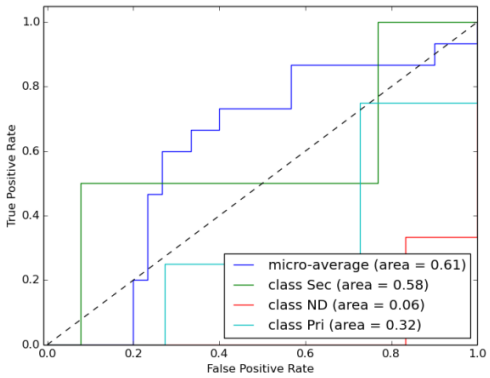
True Labels



Disease	Specificity	Sensitivity
Primary	1 (CI: 0.99, 1.0)	1 (CI: 0.99, 1.0)
Secondary	0.89 (CI: 0.86, 0.93)	0.8 (CI: 0.73, 0.86)
ND	0.9 (CI: 0.86, 0.93)	0.8 (CI: 0.72, 0.85)

Overall Accuracy: 0.86 (CI: 0.83, 0.89)

Randomly Shuffled Labels



Disease	Specificity	Sensitivity
Primary	0.97 (CI: 0.95, 0.99)	0.02 (CI: 0.01, 0.05)
Secondary	0.2 (CI: 0.15, 0.25)	0.7 (CI: 0.66, 0.80)
ND	0.75 (CI: 0.71, 0.80)	0.1 (CI: 0.05, 0.15)

Overall Accuracy: 0.32 (CI: 0.25, 0.31)

Figure 3.3: Primary vs. Secondary SVM ROC curve for real labels and shuffled labels – A multiclass SVM was trained and tested with three classes: Primary dengue infection (Pri), secondary dengue infection (Sec) and Non-Infected Donor (ND). The left side ROC indicates real labels and predicts performance of the model on new samples. The right side is the SVM tested on randomly shuffled class labels, and should be unpredictive. These curves show there is a real difference between random labels and true labels, indicating the model is not over-fitted.

Classification Summary and Dengue Subtype Prediction

Each classification experiment yielded good performance relative to randomly shuffled labels under the current scheme. This performance may also extend to subtypes of Dengue. Unfortunately, only one sample of each subtype (Dengue 1-4) could be procured from the World Health Organization making a classification experiment impossible. Though the sample size is small, single sample T-Tests picked significant peptides that survive strict (Bonferroni) multiple testing correction **(Table 3.4)** in all samples except for Dengue 3. There were eight peptides that distinguish Dengue 1 from all other subtypes and non-infected, three that distinguish Dengue 2 from others, and one that distinguishes Dengue 4 from others. This is a modest number, but the sample sizes were extremely small and the multiple testing correction strict. The other classification tests used an FDR < 0.05 cutoff, and the number of features selected varied depending on the test. In the case of the binary scheme (Dengue vs. Non-Infected) 352 peptides were selected in the IgG assay and 816 were selected in the IgM assay. There were only 28 peptides common between the two, indicating no significant overlap between the informative peptides from IgG versus IgM (P=0.79).

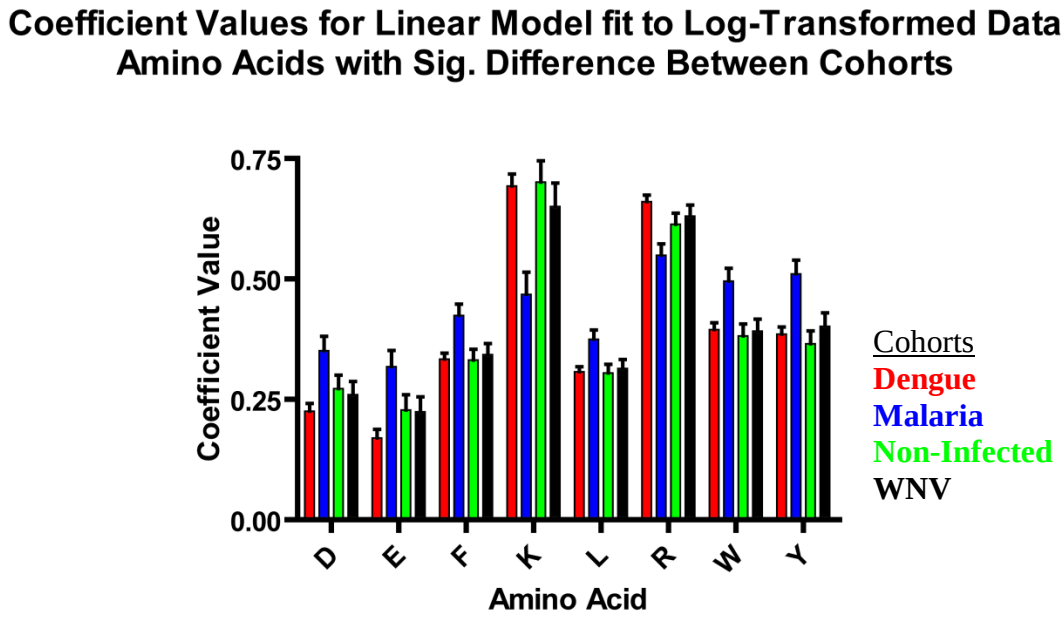
Dengue Duo IgM and IgG Capture ELISA

Sample	Country	Panbio ELISA IgG	Panbio ELISA IgM	Prediction
PVD201-10	Ecuador	0.4	5.8	Primary
PVD201-16	Ecuador	0.7	5.8	Primary
PVD201-20	Ecuador	0.7	5.8	Primary
PVD201-01	Colombia	2.4	5.8	Secondary
PVD201-02	Honduras	1.2	0.2	Secondary
PVD201-03	Honduras	1.3	1.1	Secondary
PVD201-05	Honduras	4.1	0.8	Secondary
PVD201-06	Honduras	4.5	0.7	Secondary
PVD201-07	Colombia	7	5.8	Secondary
PVD201-08	Honduras	2.6	0.9	Secondary
PVD201-13	Honduras	1.3	0.3	Secondary
PVD201-15	Honduras	1.3	0.4	Secondary
PVD201-17	Colombia	2.3	5.8	Secondary
PVD201-18	Honduras	2.3	0.3	Secondary
PVD201-21	Ecuador	4.3	5.8	Secondary
PVD201-04	Honduras	0.5	0.4	False Negative
PVD201-09	Honduras	0.7	0.3	False Negative
PVD201-11	Honduras	0.5	0.2	False Negative
PVD201-12	Honduras	0.9	0.2	False Negative
PVD201-14	USA	0	0	False Negative
PVD201-19	Honduras	0.6	0.8	False Negative

Table 3.4: ELISA Results for Predicting Primary and Secondary Infection: The Dengue Duo IgM and IgG capture ELISA is a technique for diagnosing dengue infections and differentiating primary from secondary infections. In primary infections, the IgM response is high, while the IgG response is low. In secondary infections, the IgG response is also high, and depending on the course of infection, the IgM may be low. Three samples were identified as having a low IgG response. These were predicted primary and used for classification.

Cohort Specific Amino Acid Bias

These results support the idea that immunosignatures could serve as a platform for classifying at least small cohorts of multiple infectious diseases. However, very little is known about which peptides each cohort prefers. In order to determine if any amino acid biases exist between cohorts, a multiple linear regression model was fit to an amino acid count matrix as explained in (Greiff et al., 2012). The coefficients on this model tended to cluster in a cohort specific manner, as evidenced by a principal component plot on the model coefficient weightings **(Figure 3.4)**. This was true particularly for malaria, which showed decreased binding to cationic amino acids and increased binding to anionic amino acids.



Principal Component Plot of Amino Acid Linear Model Coefficients

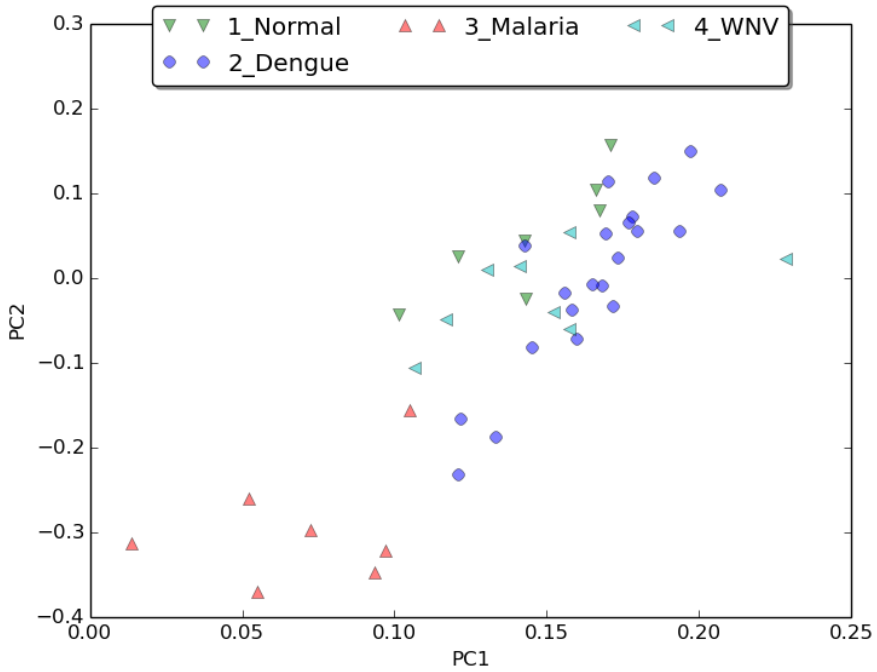


Figure 3.4: Amino Acid Linear Model Coefficients show Cohort Specific Variation: A linear model based on amino acid counts for each peptide was fit to the multiclass log-transformed and normalized data. Some amino acids show clear biases toward a particular cohort. Top shows all amino acids that had significant differences in their fitted coefficients between cohorts. Malaria especially has differential binding to several amino acids. Top shows a principal component plot of the same coefficients, showing cohort

specific variation along several amino acid axes. Samples differ primarily along the Arginine/Lysine and Aspartic/Glutamic acid axes.

Epitope Determination on 330K Arrays (IgG)

A subset of the dengue and non-infected samples were run on an array (HT330K) containing over 330,000 sequences and 27% of all possible 5-mers within its peptides (Legutki et al., 2014). These arrays engender the possibility of mapping epitopes to Dengue. Enrichment analysis revealed significant subsequences ($P < 0.005$) in seven of the twenty one dengue samples tested, with the conserved sequence AVH found in three samples (PVD201-07, PVD201-08, PVD201-16). Previous experiments (Mol. Cell Proteomics, Submitted) have shown this conserved sequence corresponds to the linear epitope AVHAD in the NS1 protein of all four strains of the virus (Y. Chen et al., 2010). Another likely epitope DANXK was also found, but it does not clearly map to the virus and has not been reported in the literature. These results are summarized in **Figure 3.5.**

Inter-alignment	18.1	14.2
Number significant	7	3750
Sample ID	PVD201-03	PVD201-04

Inter-alignment	15.1	16.6
Number significant	17	3
Sample ID	PVD201-06	PVD201-07

Inter-alignment	19.8	13.7
Number significant	204	1138
Sample ID	PVD201-08	PVD201-16

Inter-alignment	13.8
Number significant	2156
Sample ID	PVD201-20

Figure 3.5: Consensus Sequence Determination on HT330K Arrays: Enrichment analysis was performed on a per-array basis for subsequences of length 5 or greater. Seven of the twenty one Dengue samples tested contained statistically significant subsequences (FDR < 0.005). Three samples (PVD201-7, PVD201-16, PVD201-20) contained the consensus sequence AVH. Previous experiments done in our group have shown this to correspond to the linear epitope AVHAD on the Dengue (1-4) NS1 protein (Richer et al, manuscript accepted, Y. Chen et al., 2010). Those found in PVD201-04 and 201-06 did not seem to represent a consensus motif but showed a clear bias for the C-terminal residues and GSG linker. PVD201-03 revealed what appears to be an epitope but does not map convincingly to the Dengue virus. Inter-alignment refers to the mean gapless alignment score by the Smith Waterman algorithm.

Discussion

The PVD201 Dengue panel was evaluated for performance on the diagnostic CIM10K arrays under several different classification conditions. First, single class IgG and IgM classification performance was evaluated against a panel of non-infected sera. Next, multiclass IgG performance was assessed with the addition of West Nile and Malaria cohorts. Then, using predictions obtained from PanBio ELISA results, the platform was evaluated for classification between primary and secondary infections. In each case, SVM "leave one from each class out" classification was performed and specificity and sensitivity were estimated with a bootstrapping procedure. Subtype classification could not be preformed due to lack of samples, but the WHO panel of Dengue subtypes was evaluated for statistically significant differences between subtypes. Additionally, the samples from each cohort were individually fit to a linear model based on amino acid counts similar to what was done by Greiff and colleagues (Greiff et al., 2012). Finally, the dengue samples and some of the non-infected samples were run on the new CIM330K arrays (Legutki et al., 2014) and dengue-specific consensus sequences were identified.

Classification results were encouraging considering this is an early stage diagnostic and existing, approved diagnostics are generally cross reactive. The PanbioIgM/IgG test employed here for example produced 0.28 false negative rate and is known to be cross reactive to other flaviviruses and malaria (Banoo et al., 2008; Hunsperger et al., 2009; Schwartz et al., 2000). The FDA approved InBios Detect IgM Capture ELISA reports 0.94 specificity and 0.92 sensitivity (Namekar et al., 2013), but the manufacturer acknowledges cross-reactivity with other flaviviruses and warns users

about this fact ("DENV DetectTM IgM CAPTURE ELISA," 2012). None of the ELISA based methods is capable of differentiating Dengue strains. The peptide array matched these results (sensitivity: 0.96, specificity: 0.94) in the single class test and also achieved 0.76 specificity and 0.96 sensitivity in a classification problem including West Nile Virus and Malaria samples. While this specificity leaves something to be desired, it is an improvement over existing techniques and with additional samples could be refined from its current experimental iteration into a rigorous and accurate diagnostic. It is a unique feature of immunosignatures that new data has the potential to improve the diagnostic.

Distinguishing primary infection from secondary was an important result from this study. It is puzzling, however, that primary infections were classified with high sensitivity and specificity (1.0, 1.0) in spite of the low IgG ELISA (**Table 3.3**). As an IgG based assay, immunosignatures should not pick up the IgM/IgG ratio differences that characterize primary versus secondary infections. A possible explanation is that the sensitivity of the arrays detects an IgG response not evident in the InBios ELISA, which uses recombinant proteins. Another, less likely but interesting possibility is that there is another IgG response to pathogen infection that does not include specific antibodies generated against pathogen targets. In other words, there may be an innate component to the IgG response that occurs in primary infections before mature B-cells appear.

The amino acid model revealed consistent binding across cohorts to each residue, with the exception of charged residues. These varied significantly for the Malaria cohort which showed increased binding to anionic residues relative to the other cohorts. This could reflect the extraordinary antigenic variation possessed by *P. falciparum* (Anders, 1986; Pasternak & Dzikowski, 2009; Scherf et al., 2008). In contrast to the other

diseases, *P. falciparum* and the other Malaria causing protists have a family of var genes that provide antigenic diversity in order to evade the host immune system (Gardner et al., 2002). Perhaps this host-pathogen interaction produces this charge differential. Alternatively, it could be caused by fluctuations in pH within the sample solutions. The latter seems unlikely considering these samples were highly diluted (1:500) into the same buffer solution before being applied to the array, and the arrays would need to be highly sensitive in order to pick up any pH differences.

An application of this technology in addition to diagnostics is epitope mapping. This is an area of active research in our group, and well designed experiments can reveal exact pentamer sequences corresponding to pathogen epitopes (Mol. Cell Proteomics, Submitted). The present study contains an early iteration of the 330K arrays which revealed a consensus trimer to a known linear, immunodominant NS1 epitope (Y. Chen et al., 2010). This partially validates an underlying assumption of this assay, which is that different infected immune systems often produce antibodies to the exact same or similar sequence on the pathogen. More data on this subject would be of great utility for refining feature selection methodology and improving array design as a diagnostic platform.

This study had two major limitations: small cohorts and non-endemic non-infected samples. It is true that additional sample variation in the form of larger cohorts could dramatically change the SVM model and classification rules, but the technology is adaptable in that more information becomes available about how dengue serum reacts to the peptide arrays. Thus, each sample run not only produces a diagnostic result, but also an additional sample which could be used to improve the test. If new subtypes are discovered, the same arrays could be used to detect this. The second limitation arises

from the use of non-endemic non-infected samples. This can affect classification results, but in this case there was an endemic Malaria cohort included which was successfully separated from Dengue.

There is a growing list of examples of immunosignatures classifying a diverse set of target disease states from non-infected or other clinically relevant diseases (Hughes et al., 2012; Kukreja et al., 2012; Legutki et al., 2014; Restrepo et al., 2011a; Sykes et al., 2013). This technology is maturing and deserves more attention as a diagnostic platform. We still have a poor understanding of why this technique works so well for such a wide variety of diseases, and which sequences are driving these results. Improved modeling techniques for relating sequence information to feature intensity would be essential for understanding the mechanisms underlying immunosignatures.

CHAPTER 4

A SIMPLE KINETIC MODEL FOR IMMUNOSIGNATURES

Abstract

High throughput assays involving multiple receptors and multiple ligands are common in modern biology. The most prominent examples of these are expression arrays, protein arrays, and peptide arrays. While these are useful assays, users often assume their kinetic behavior is similar to what one would expect in an experiment using a single receptor. This chapter challenges this assumption, developing a simple first order kinetic model for multiple-receptor multiple-ligand kinetics. This model is developed and tested in the context of the immunosignature assay: non-natural sequence peptide microarrays designed to measure and quantify circulating antibodies in blood or sera. The model unveils several surprising kinetic phenomena. According to the model, these assays are highly sensitive to context: each individual ligand-receptor interaction is sensitive to all the other receptors present on the array. There is also a critical balance between total receptor (peptide) and total ligand (antibody) concentration. Finally, there is a very interesting crossover effect observed under certain assay conditions, whereby ligand-receptor association curves are non-monotonic (they initially increase, then decrease) due to context-dependant competition. This result is counter to behavior under single receptor conditions, but has been verified in experiments involving antibody-peptide interactions on peptide microarrays. This work explains a number of previously poorly understood and unexpected results, and paves the way for further experiments aimed at optimizing high throughput platforms.

Introduction

Peptide arrays have proven useful as a method for finding high affinity ligands (Diehnelt et al., 2010), characterization of monoclonal and polyclonal antibodies (Halperin et al., 2011) and as a diagnostic method using immunosignatures (Halperin et al., 2011; Legutki et al., 2010; Restrepo et al., 2012; Restrepo et al., 2011b; Sykes et al., 2013). These assays may involve ~10^9 antibody species at unknown concentrations in addition to a deliberately complex peptide coated surface (Legutki et al., 2014). Due to this complexity, the biophysics behind peptide-antibody binding in peptide array based assays is poorly understood (Stafford et al., 2012). There are reams of (unpublished) experimental data showing that immunosignatures are highly dependent on assay conditions in surprising ways, indicating a better kinetic model is needed to explain and design appropriate assay conditions. This work reviews the surprising findings from previously collected immunosignature data, defines a kinetic model of a peptide microarray assay which explains these findings, and tests the model against some recently collected data.

An immunosignature assay involves screening a large number of peptides against serum samples, then picking a subset of "informative" peptides for use as a diagnostic. A natural idea is to create a second array containing only this informative subset in order to improve reproducibility, ease analysis, and generally optimize the method. This was attempted in The Center for Innovations in Medicine (CIM) by several people (Bart Legutki, Krupa Navalkar, Xiao Wang and others), and these efforts all concluded that informative peptides resulting in good classification in the large array did not result in as good a classification when the sub-array was used, despite the fact that they contained the

same peptides and were run under the same assay conditions (Navalkar et al., 2014). Closer analysis revealed that peptide intensities from the sub-array were less correlated with their corresponding peptides from the larger array (unpublished). These experiments showed context-sensitivity indicating a kinetic aspect to these assays that is as-yet poorly understood.

A second experiment involved the use of PepPerPrint arrays (**Figure 4.1)**, which were manufactured using a different method than those used previously in CIM. These arrays are produced using a proprietary process and are optimized for epitope mapping. Due to this requirement, the density (number of peptides per unit area per feature/spot) could be low in order to minimize cross-reactivity with mimotopes (but actual density is unknown). Peptide arrays from CIM (CIM10K) are printed on aminosilane glass slides (Legutki et al., 2010), and seek to maximize density which seems to result in an improved ability to measure low affinity antibody-peptide interactions. The same sera and monoclonal antibodies were applied to these two different arrays under the same assay conditions, and binding to surface peptide was detected using a fluorescently labeled secondary antibody. Despite using the same assay conditions and having some of the same peptide subsequences, almost no fluorescence was observed using the PepPerPrint arrays compared to the CIM10K either using serum or using monoclonal antibody as the ligand. This suggests a fundamental difference between arrays printed on aminosilane and those synthesized by PepPerPrint, most likely related to the density of each peptide within each spot.

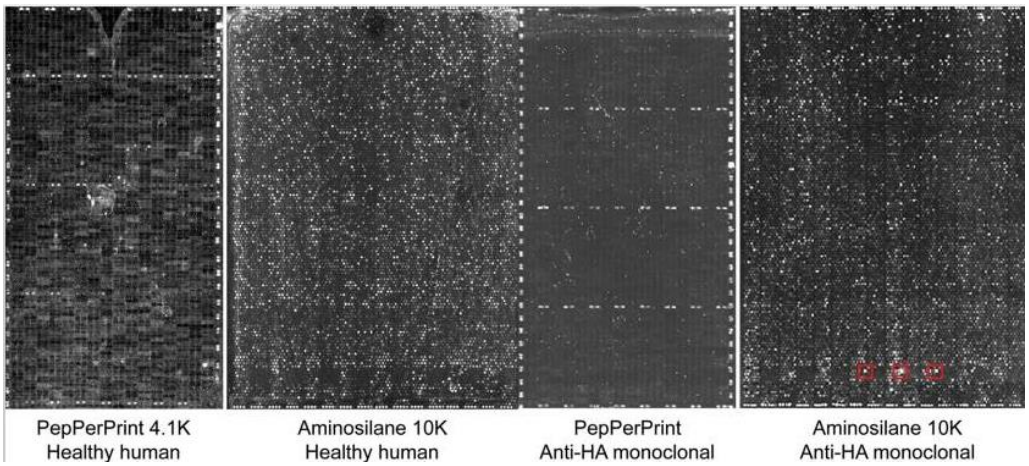


Figure 4.1: PepPerPrint Arrays vs. CIM10K: Reproduced with permission from Sykes et. al. (Sykes, Legutki, & Stafford, 2013). This figure compares the same samples run on PepPerPrint arrays versus aminosilane slides. PepPerPrint arrays are inferred to be much less dense (fewer peptides per unit area) than the aminosilane slides.

An additional experiment (Navalkar, 2014) involved an array containing almost 5000 peptides from various pathogen proteomes. Two versions of this array were created: one including influenza peptides and another leaving them out. This consisted of around 1% of the total peptides on the array, and the others were kept constant between the two versions. The addition of influenza peptides greatly changed the classification results and the pattern of binding. Further analysis revealed that the arrays without influenza peptides had a higher average fluorescence than those containing the influenza peptides **(Figure 4.2)**. It appeared as though influenza peptides were "stealing" reactivity from the other peptides. This effect was particularly consistent among the low intensity peptide-antibody complexes.

Another important point about these assays is that most antibody-peptide interactions on these arrays are low affinity (micromolar range) in solution (data not shown). This is expected since these peptides are random and most have no relationship to an epitope to a given antibody. There is a density effect that needs to be explained,

allowing us to observe low affinity interactions even at very small concentrations of antibodies. Some hypothesize that high peptide density at the surface artificially increases this affinity. Others believe that surface affinities are similar to those measured in solution, and the density effect can be explained by simple thermodynamic laws related to the total concentration of peptide (number of binding sites) being so large such that the reaction is driven towards antibody-peptide complex formation. The model generated here tests both possibilities.

These results and observations constitute a theme that has affected immunosignatures from the start, which is that peptide context and assay conditions matter in unexpected ways. If immunosignatures are to become a reliable and robust diagnostic method, a better understanding of the kinetic and biophysical aspects of the assay is needed. To this end the following sections review classical receptor-ligand kinetics and develop a model for peptide-antibody interactions on a peptide array. This model is compared to the existing data available and experiments are proposed that would test the model.

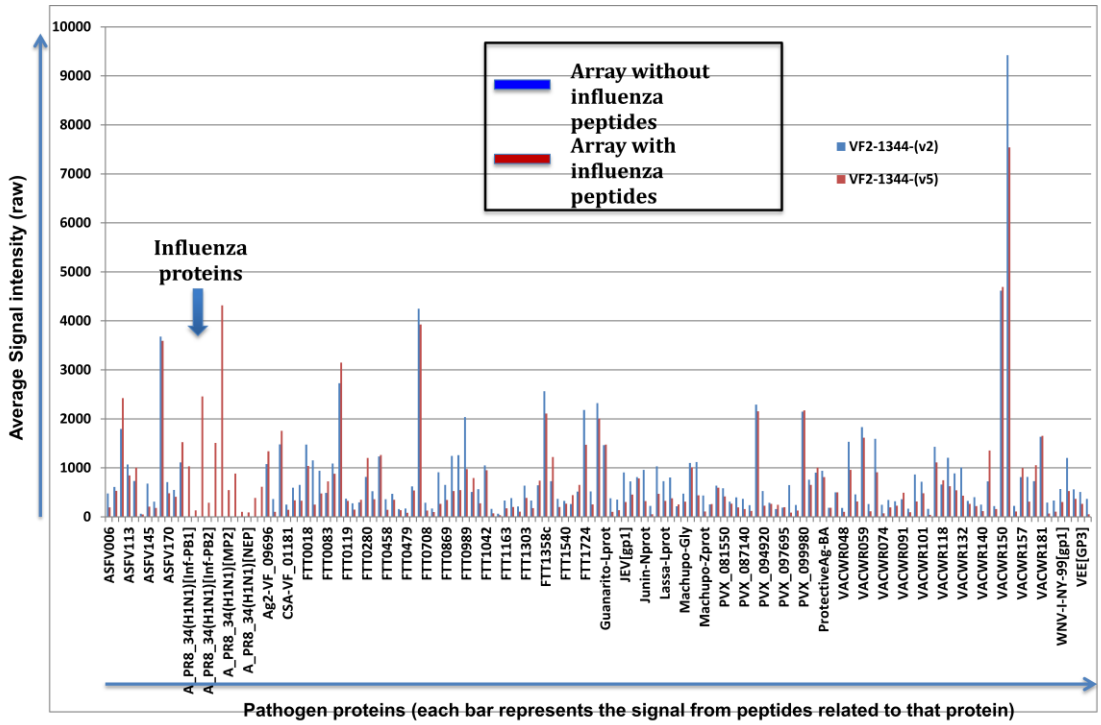


Figure 4.2: Effect of Additional Peptides on an Array: Reproduced with permission from Krupa Navalkar. This experiment tested two different arrays against the same biological sample. Arrays with influenza peptides added yielded lower overall signal for the rest of the peptides on the array as compared to the array without influenza peptides. This indicates a context dependence whereby the final signal of any given peptide depends on all the other peptides on the array.

Modeling Receptor-Ligand Kinetics

Single Receptor with a Single Ligand

Consider a system consisting of a single receptor species and a single ligand species in solution.

$$R + L \leftrightarrow RL$$

Such a system has well a well defined dissociation constant (affinity) which depends on two kinetic constants k_{on} and k_{off}.

$$K_d = \frac{k_{off}}{k_{on}} = \frac{[R]_{eq}[L]_{eq}}{[RL]_{eq}}$$

The above assumes an equilibrium has been reached between R and L.

The first order kinetic rate equations for R, L and RL are

$$\frac{d[R]}{dt} = -k_{on}[R][L] + k_{off}[RL]$$

$$\frac{d[L]}{dt} = -k_{on}[R][L] + k_{off}[RL]$$

$$\frac{d[RL]}{dt} = k_{on}[R][L] - k_{off}[RL] \quad (1)$$

Since the three species are interdependent and can be expressed in terms of constants and one of the species, this system is separable, and has an analytical solution. The equation for predicting $[\mathrm{RL}]$ at time t is:

$$RL(t) = \frac{k_{on}[L_0][R_{tot}]}{k_{on}[L_0]+k_{off}} \times (1 - e^{-(k_{on}[L_0]+k_{off})t}) \qquad (2)$$

This treatment is common in kinetic assays such as SPR, and is the common mental model of practitioners of kinetic assays. The ligand is added to excess such that it is not significantly depleted during the assay, and equation (2) is then fit to experimental data in order to determine kinetic rates. immunosignatures are a far more complicated situation and it will be shown that this regime does not directly translate to assays involving multiple receptors and multiple ligands.

Multiple Receptors with a Single Ligand

Consider a system consisting of *n* peptides (receptor) and one ligand (antibody). This is the situation with monoclonal antibodies on an array (assuming each peptide binds an antibody only one way).

$$L + R_1 + R_2 + ... + R_n \leftrightarrow LR_1 + LR_2 + ... + LR_n$$

The rate equations for this system are as follows:

$$\frac{d[R_i]}{dt} = k_{off}^{(i)}[LR_i] - k_{on}^{(i)}[L][R_i]$$

$$\frac{d[L]}{dt} = \sum_i k_{off}^{(i)}[LR_i] - \sum_i k_{on}^{(i)}[L][R_i]$$

$$\frac{d[LR_i]}{dt} = -k_{off}^{(i)}[LR_i] + k_{on}^{(i)}[L][R_i]$$

This is the situation when monoclonal antibodies are applied to an array containing *n* peptides. This is closer to reality than the single receptor-single ligand model, but it can be further generalized to include the possibility of multiple ligands as well as multiple receptors.

Multiple Receptors with Multiple Ligands

Finally, we come to the situation in immunosignatures, where multiple antibodies are interacting with multiple peptides in competition. This formulation was inspired by work done by Dr. Neal Woodbury (unpublished).

$$L_1 + ... + L_m + R_1 + ... + R_n \leftrightarrow L_1R_1 + L_1R_2 + ... + L_1R_n + ... + L_mR_n$$

The rate equations underlying this system are:

$$\frac{d[R]_i}{dt} = \sum_j k_{off}^{(i,j)}[L_jR_i] - \sum_j k_{on}^{(i,j)}[L_j][R_i] \qquad (3)$$

$$\frac{d[L_j]}{dt} = \sum_i k_{off}^{(i,j)}[L_jR_i] - \sum_i k_{on}^{(i,j)}[L_j][R_i] \qquad (4)$$

$$\frac{d[L_jR_i]}{dt} = -k_{off}^{(i,j)}[L_jR_i] + k_{on}^{(i,j)}[L_j][R_i] \qquad (5)$$

Surface Considerations

The above equations assume all species are in a well mixed solution. In reality the receptors are peptides on a surface, fixed at the end of the linear sequence (N or C terminus depending on the manufacturing process). The most obvious concern is diffusion, which may limit the rate at which ligands can access peptides at the surface. This may artificially slow k_{on}, but a spatial concentration gradient during incubation is unlikely due to constant applied agitation, so spatial effects are ignored here. Other

effects, such as density are more likely to play a large role in binding behavior and are considered below.

Density

The density of peptide on a spot is unknown but thought to be high, possibly 1 peptide per nm^2 (Neal Woodbury, unpublished). Recall from equation 1 (and Le Chatelier's Principle), the rate of complex formation depends on both $[L]$ and $[R]$. If $[R]$ is high, this drives up the rate of complex formation and shifts equilibrium conditions toward the complex. The role of density in immunosignature assays is highly suspected due to the results of the PepPerPrint experiments **(Figure 4.1)** (Stafford et al., 2012) and experiments on NSB slides which have a known, fixed density (Navalkar, 2014).

Picking Parameters

Equations 3,4, and 5 describe kinetics of a peptide array system, with many simplifying assumptions. The model treats peptides and antibodies as if they are floating around in solutions, ignoring the surface entirely. Also it does not consider blocking competition, which may dampen any low affinity binding effects. Still, it may be a useful model for understanding how changes in assay conditions affect observed results.

The question now becomes how to parameterize the model. Here due to lack of some important data, some assumptions can be made. First, we are interested primarily in how assay conditions and relative incubation times change the binding profile (e.g. what happens when you change assay volume or double incubation time). Since the actual values of K_{on} and K_{off} are unknown, we take a "best guess" approach and focus on the ratio K_d for which we have a better idea of what is expected. This means that estimating time to equilibrium is beyond the scope of this model as parameterized here, but this is

not an important factor for the questions being evaluated. We assume K_{on} are constant at $0.001s^{-1}$ and need to generate a matrix of K_{off} values such that K_d*'s* fall within an expected range.

In solution, this range likely is loaded heavily toward the high micromolar range with some small number of nanomolar affinities. The distribution is unknown, and would ideally be obtained from detailed measurements which do not yet exist. Later we discuss approaches for obtaining these detailed measurements. For now, we simply assume the K_d values are Pareto distributed. This distribution was named after the Pareto, or "80/20" rule. Though the actual proportions vary based on its parameter α, intuitively this means that 80% of antibodies are low affinity and 20% are high affinity.

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_\mathrm{m}}{x}\right)^\alpha & x \geq x_\mathrm{m}, \\ 0 & x < x_\mathrm{m}. \end{cases}$$

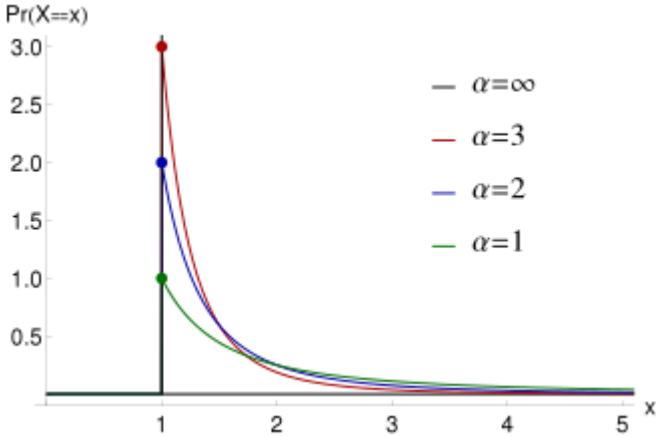


Figure 4.3: Pareto Distribution and Properties: The Pareto distribution has two parameters: α and x_m. x_m specifies a lower bound on the distribution while alpha dictates the shape of the curve. We are interested in generating micromolar to nanomolar affinities between 1 nM and 1 mM, with loadings in the uM range. To generate these values, this distribution must be modified such that generated values fall within a user-defined range. Figure reproduced from Wikipedia (Public Domain).

We generate the K_{off} matrix assuring that K_d values fall within a user-set range. This is derived as follows:

$$X_{i,j} \sim Pareto(0, \infty)$$

$$K_d^{(i,j)} = \frac{X_{i,j} - min(X)}{max(X)} * (h - l) + l \qquad (6)$$

Equation 6 is a function of three arguments which generates a matrix of K_d values:

$$K_d = f(X, l, h) \qquad (7)$$

Where the shape of K_d is the same shape as the input matrix X of Pareto distributed values, and where *l* and *h* are the low and high bounds of K_d in the distribution (say, between 100nM and 1M). Then the actual k_{off} values used in the simulation can be computed easily:

$$k_{off}^{(i,j)} = K_d^{i,j} * k_{on} \qquad (8)$$

This generates a matrix of off rates which can be use to parameterize equations 3, 4, and 5. Now the initial concentrations of peptide and antibody can be set to mimic different assay conditions. The initial conditions tested are designed to evaluate the following situations:

- The effect of volume
    - In this model, increased volume would decrease the effective peptide concentration, since the number of peptides on the surface remains fixed. Thus, to simulate a doubling of volume we simply halve the initial free concentration of each peptide. How does this affect the binding profile with all else remaining constant?

- Ratio of peptide concentration to antibody concentration
    - This is related to the volume effect. There are three possible **regimes** here:
        - **Regime A**: $[P] > [A]$ antibody is the limiting reagent
        - **Regime B**: $[P] = [A]$ antibody and peptide are roughly equal
        - **Regime C**: $[P] < [A]$ peptide is the limiting reagent

- As volume changes, $[P]$ decreases, meaning that if immunosignatures work in regime A or B, the assay could be volume sensitive in that changes could cause it to fall into the middle or bottom regime.

Method: Volume Experiment

Several human serum samples were tested on the same arrays containing over 10,000 non-natural sequence peptides (CIM10Kv3). These were tested under standard dual-channel (IgG, IgM) experimental conditions (see chapter 3 on Dengue Diagnostics). The only experimental parameter varied between replicates was the solution volume containing the primary antibody (the serum), which was varied at 240uL and 600uL. Primary antibody dilution was 1:1000 corresponding to approximately 66nM (assuming sera contains ~10mg/ml antibody). The objective was to test the sensitivity of array results to total volume, which these assays normally would be insensitive to under traditional modes of thinking about assays like these. These experiments were conducted by Phillip Stafford and the Peptide Array Core.

Method: Incubation Time Experiment

25nM of monoclonal antibody (p53Ab1) was applied to the CIM10Kv2 array containing 10,000 non-natural sequence peptides. This was run under standard assay conditions (Dengue Diagnostics, Chapter 3, Halperin et. al. (Halperin et al., 2011)) with the exception that the incubation time in one of the conditions was increased to 16 hours.

Method: Kinetic Simulations

Simulations of the kinetic system given in equations 3,4, and 5 was conducted using Euler's method. The code for accomplishing this is hosted at:

https://github.com/joshuaar/CIM-Scripts/blob/master/pepKineticModel.py

It has a simple command line interface for varying parameters.

Results

Effect of Peptide Dilution Under the Model: A Testable Prediction

First I designed a testable prediction under the model and evaluate it against experimental data. One such prediction is to observe the effect of peptide dilution while keeping antibody concentration fixed. In a real experiment, one could achieve this by varying the volume of the primary antibody solution (which effectively dilutes the peptides at the surface), while keeping antibody concentration fixed. Under the model, the distribution of equilibrium complex loadings for each peptide should be sensitive to the ratio of total antibody to total peptide unless one of the two is present to sufficient excess. I tested 5 different peptide concentrations relative to a fixed 5nM antibody concentration. Each condition used 200 peptides, 20 antibodies, and K_d values ranging from 10 to 100nM.

The choice of K_d here is arbitrary, as this only affects the concentration ranges at which the equilibrium distribution is sensitive to change. After simulating these conditions under the model, the variance of the peptide loadings at equilibrium was calculated. This can be thought of as a virtual array experiment, where 20 antibodies are assayed against 200 peptides, the reaction is stopped at equilibrium and complex concentrations are detected via fluorescence. These conditions and variance results are given in **Table 4.1**. The distributions at equilibrium are visualized in **Figure 4.4**.

[Ptotal] (nM)	5000	500	50	5	0.5
[Abtotal] (nM)	5	5	5	5	5
N Ab	20	20	20	20	20
N Peptide	200	200	200	200	200
Variance	1.03×10^{-10}	3.52×10^{-9}	3.32×10^{-7}	2.15×10^{-6}	1.55×10^{-6}

Table 4.1: Variance Results for Peptide Dilutions: These are the equilibrium variances for each experimental condition tested under the model. There is a clear trend of decreasing variance as peptide concentration decrease relative to antibody concentration.

As shown earlier, this also corresponds to lower overall binding. While the lower overall binding could be compensated for by better scanning lasers and PMT settings, the variance trends cannot be changed through better imaging techniques.

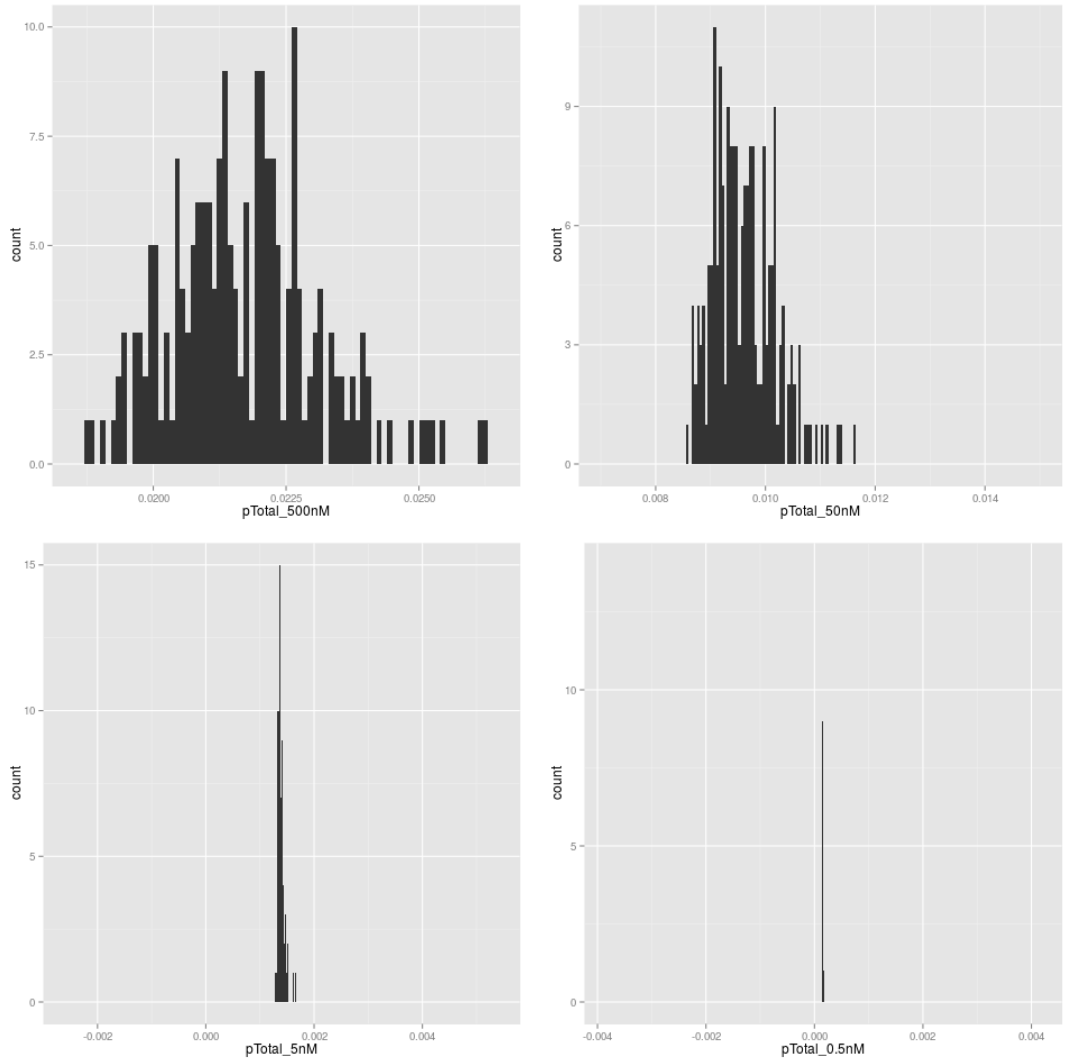


Figure 4.4: Generated Distributional Trends Under the Kinetic Model: Under the model, the variance in equilibrium peptide loadings decreases as peptide concentration is decreased relative to antibody concentration. Top left: 500nM peptide, 5nM antibody. Top right: 50nM peptide, 5nM antibody. Bottom left: 5nM peptide, 5nM antibody. Bottom right: 0.5nM peptide 5nM antobody. All x-axis scales have the same width but different locations, and refer to the distribution of equilibrium peptide-antibody complex concentrations for the 200 tested peptides.This provides a hypothesis that can be tested experimentally. If immunosignature experiments exist in regime A, they should be volume sensitive. These distributions were generated under the model allowing all complexes to reach equilibrium. 20 antibodies with a total concentration of 20 nM were

run against 200 peptides with total concentrations ranging from 5000 to 0.5 nM. Based on these results, variance becomes insensitive once peptide is sufficiently concentrated (5000nM and 500nM distributions are nearly identical, 5000nM not shown).

The simulations show that unless peptide is extremely in excess, the equilibrium distribution is sensitive to volume of the primary antibody solution (as this effectively dilutes peptide on the surface). To see if this takes place in reality, an experiment was designed on the 10Kv3 arrays and Sera.

Experimental Results: Volume Experiment

Experiments varying volume at a fixed antibody concentration effectively dilute or concentrate the effective peptide concentration. Thus, it is possible to test whether this kinetic model has bearing on reality. The experiment is simple. Place some sera on an array at a fixed concentration and incubate under standard conditions. Do several replicates, each at different volumes. If there is a significant difference in the distribution of intensity values, then current immunosignature conditions most likely exist in antibody limiting conditions (regime A). In both the IgG and IgM cases, the total variance was sensitive to volume (IgG: P=0.0024, IgM: P=0.00029), indicating that regime A is the most likely situation for immunosignatures, and the model appropriately captures this fact. These results are summarized in **Figure 4.5**.

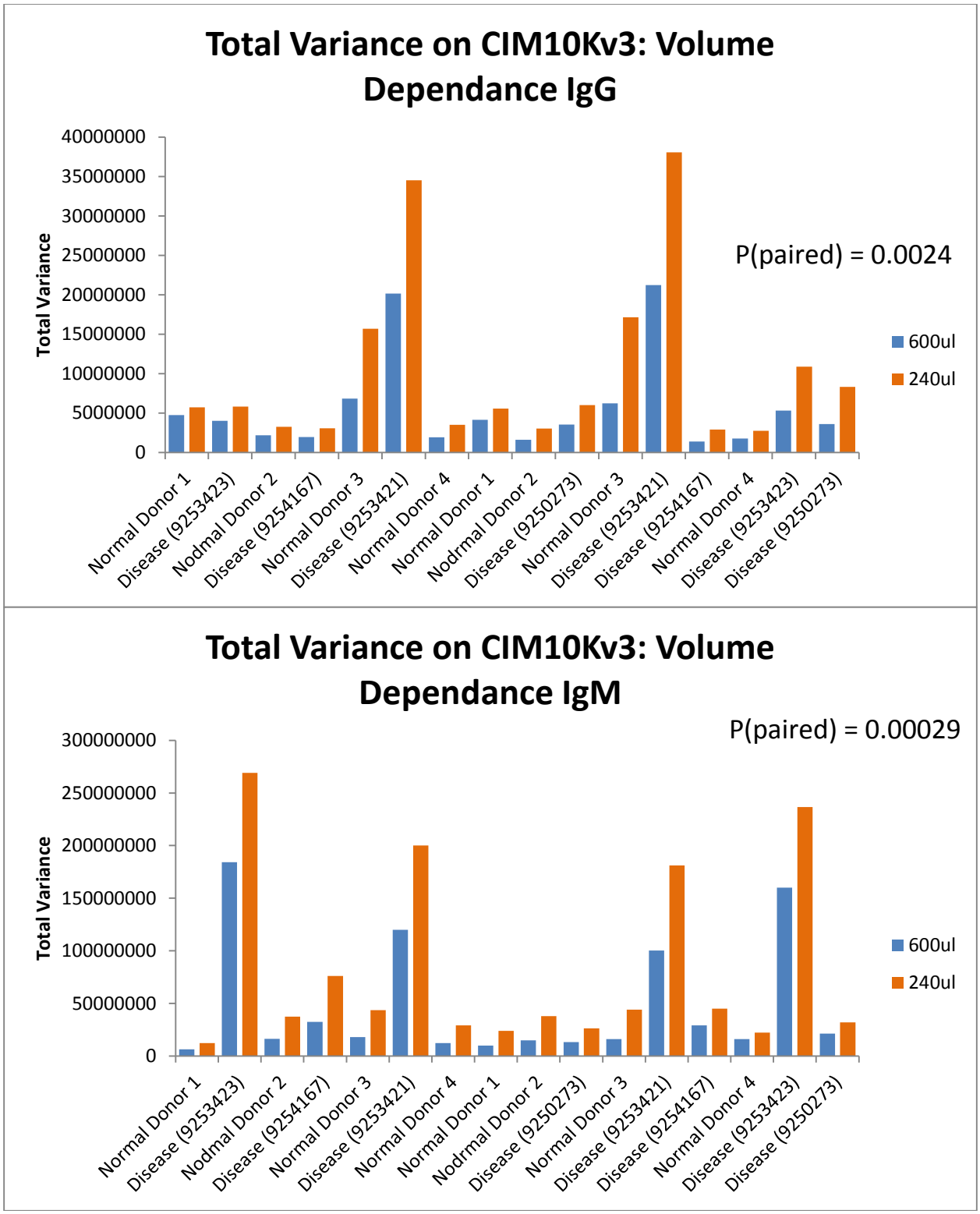


Figure 4.5: Volume Dependence on Total Variance of Array Results: The results shown here are consistent with those predicted by the model if conditions are in regime A. Higher volumes effectively dilute peptide concentrations, resulting in a lower

concentration of peptide relative to antibody. The model predicts that this would result in an overall reduction in variance, which is very clearly observed here. The fact that these peptides are stuck on a surface rather than floating around in solution does not seem to change this fundamental law, and is further evidence that immunosignatures exist in regime A.

Context Sensitivity and The Crossover Effect

There are two strange phenomena which have been observed experimentally on these arrays which I believe to be related. One is context sensitivity, which was covered in the introduction and means that the equilibrium fluorescence intensity of each peptide depends on every other peptide on the array. The second observed phenomenon is the "crossover effect" whereby some peptides show high fluorescence intensity at short incubation times, but low intensity at long incubation times. Both of these are unexpected, but the model sheds light on what conditions could give rise to these effects. An example of the crossover effect is shown in **Figure 4.6.** It will be shown that the kinetic model developed here can explain this crossover effect as well as context dependence if certain conditions are met.

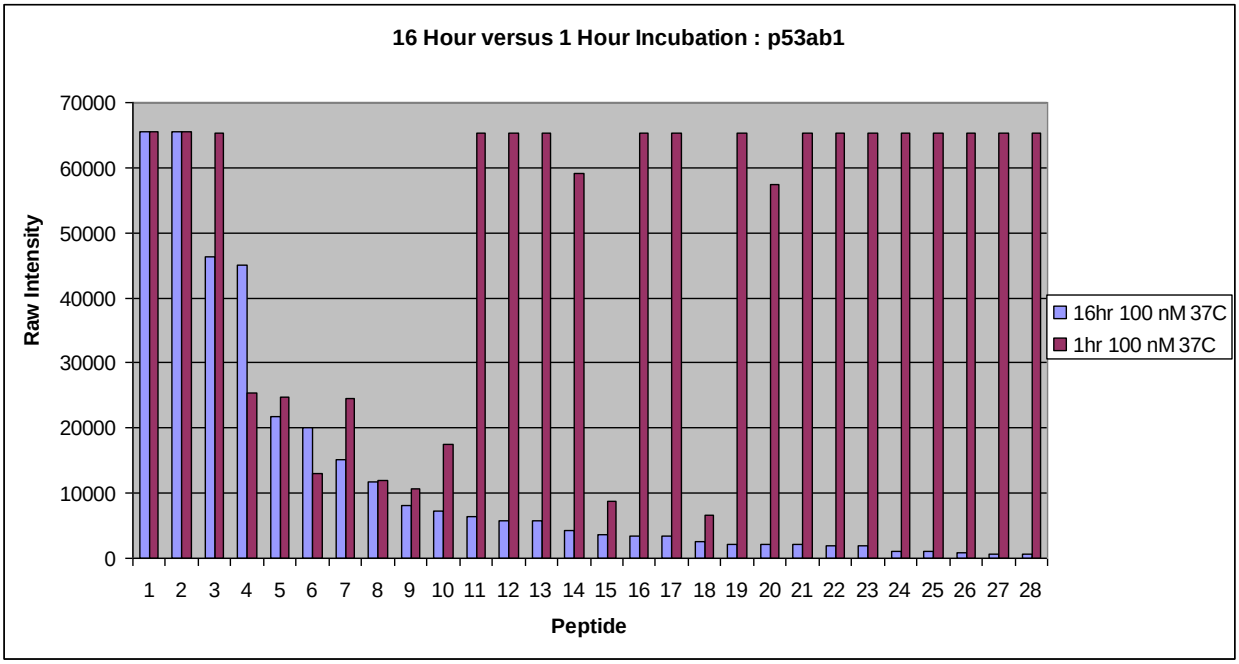


Figure 4.6: Crossover Effect in Monoclonal Antibody: The crossover effect is when initially high intensity peptides decreased in intensity as the assay is allowed to run to equilibrium. Here is an experimental result where a monoclonal antibody was incubated on the CIM10K array, and peptides were ranked by their intensities at 16 hours. It is clear that some peptides which have very low intensity at 16 hours were nearly saturated at 1 hour. The kinetic model explains these results if certain conditions are met.

Behavior Under Different K_d, Peptide and Antibody Concentrations

In order to explain context sensitivity and the crossover effect, a number of different experimental conditions were simulated under the model. These are given in **Table 4.2**. It will be shown that when peptide is sufficiently in excess (conditions 1, 2, 7, 13), crossover effects due to context sensitivity are observed. "Sufficient excess peptide" is dependent on the K_d matrix of the antibody peptide interactions, but it is always the case that given enough peptide, crossover effects and context dependence can be observed.

Several kinetic systems were tested under different conditions:

Hypothesis	Regime	Condition	N Abs	N Peptides	$[Ab_{total}]$	$[Peptide_{total}]$	K_{on}	K_{off}
X	A	1	20	20	5nM	5000nM	0.001 $nM^{-1}s^{-1}$	
	A	2	20	20	5nM	500nM	0.001 $nM^{-1}s^{-1}$	
	A	3	20	20	5nM	50nM	0.001 $nM^{-1}s^{-1}$	0.01 – 0.1 s^{-1}
	B	4	20	20	5nM	5nM	0.001 $nM^{-1}s^{-1}$	See eqn. 8
	C	5	20	20	500nM	5nM	0.001 $nM^{-1}s^{-1}$	
	A	6	1	20	5nM	500nM	0.001 $nM^{-1}s^{-1}$	
Y	A	7	20	20	5nM	50000nM	0.001 $nM^{-1}s^{-1}$	
	A	8	20	20	5nM	5000nM	0.001 $nM^{-1}s^{-1}$	
	A	9	20	20	5nM	500nM	0.001 $nM^{-1}s^{-1}$	
	A	10	20	20	5nM	50nM	0.001 $nM^{-1}s^{-1}$	1-10 s^{-1}
	B	11	20	20	5nM	5nM	0.001 $nM^{-1}s^{-1}$	See eqn. 8
	C	12	20	20	500nM	5nM	0.001 $nM^{-1}s^{-1}$	
Z	A	13	20	10	5nM	100nM	0.001 $nM^{-1}s^{-1}$	0.01 – 0.1 s^{-1}

Table 4.2: Experimental Conditions Tested: Each experimental condition is divided into one of three regimes. Regime A refers to where peptide is present in excess, regime B is where peptide and antibody are roughly equal, and regime C is where antibody appears in excess. These conditions were chosen to test the effect of these ratios on equilibrium distributions.

As mentioned previously, there are two competing hypotheses about the role of density on peptide arrays. The first states that the high density artificially decreases k_{off} by some unknown mechanism. I call this hypothesis X. The other hypothesis states that this mechanism is unnecessary, surface affinities are similar to those measured in solution, and the observed binding is simply due to excess peptide on the surface and known thermodynamic laws. We call this hypothesis Y. There is also the question of what happens when fewer peptides are represented on the array with the total peptide concentration kept constant. These situations are considered below.

Figure 4.7.1: Hypothesis X, Regime A, Condition 1: Extremely Dense Peptide, 20 antibodies with 20 peptides

This tests an absolutely extreme case, whereby there is 1000x more peptide than antibody on the array. Total antibody is completely depleted very rapidly due to the massively excess peptide concentration, after which they come off their initial targets and settle on those interactions for which they have the lowest K_d. In this range, antibody is almost totally depleted at equilibrium.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	5000nM (250nM)	10 – 100 nM

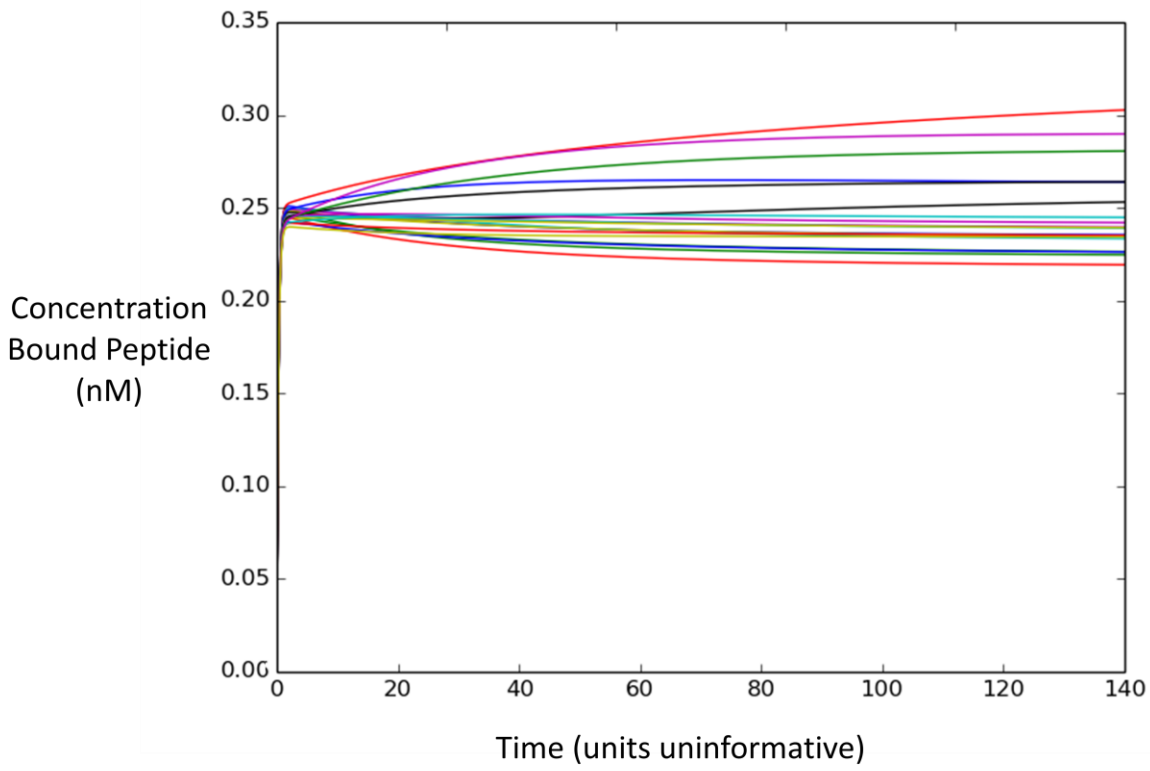


Figure 4.7.2: Hypothesis X , Regime A, Condition 2: Dense Peptide, 20 antibodies with 20 peptides

While still in regime A, in this test there is only 100x more peptide than antibody**.**

Behavior is similar to the 1000x case, but the initial depletion is slower and there is a more pronounced "crossover" effect, indicating that under these conditions it becomes very critical at which point the assay is stopped.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	500nM (25nM)	10 – 100 nM

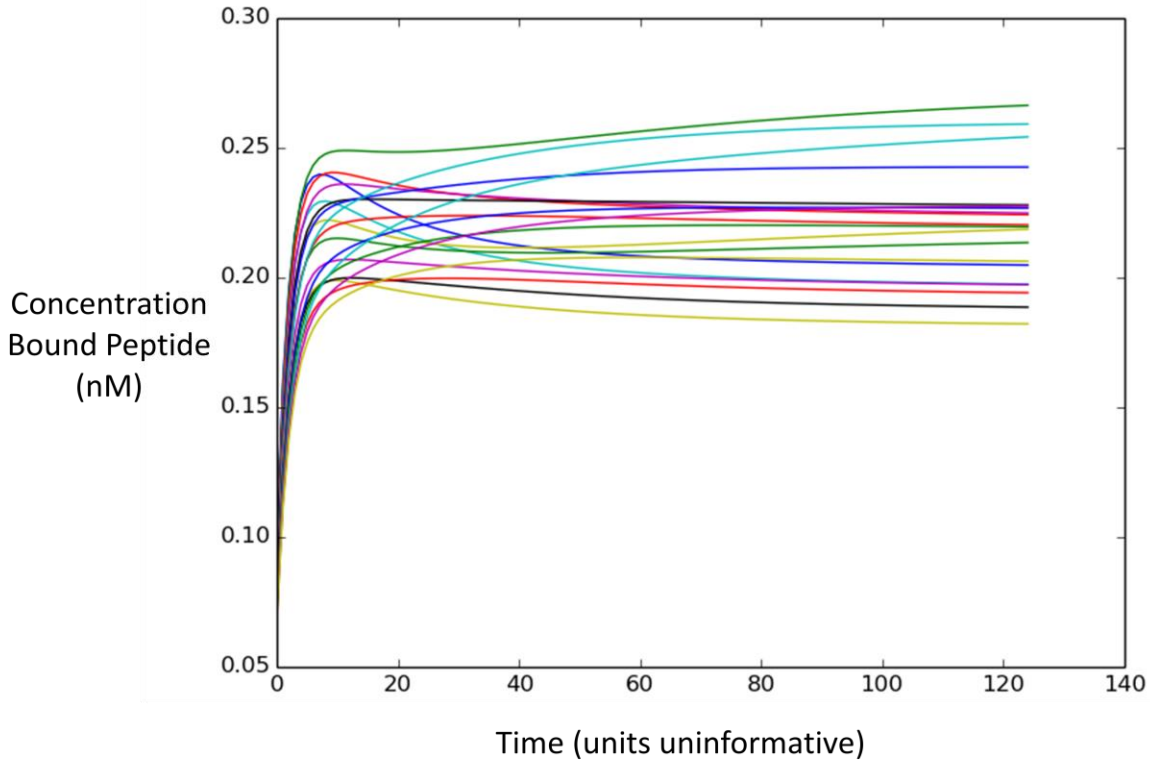


Figure 4.7.3: Hypothesis X , Regime A, Condition 3: Moderately Dense Peptide, 20 antibodies with 20 peptides

Here peptides are 10x more concentrated than antibodies. At this point, antibodies are not depleted at equilibrium, and the bound concentration varies only by around 0.02 nM. This regime is unlikely to result in a good immunosignature.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	50nM (2.5nM)	10 – 100 nM

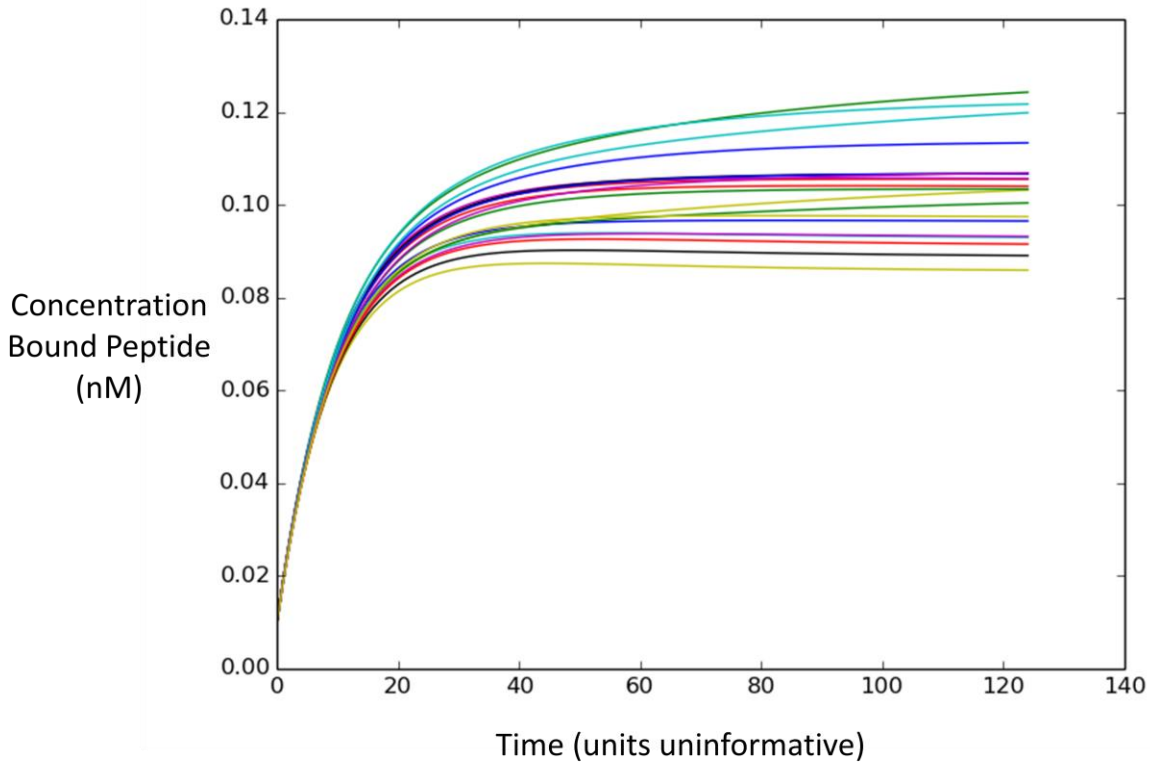


Figure 4.7.4: Hypothesis X , Regime B, Condition 4: Equality Peptide, 20 antibodies with 20 peptides

Here total peptide and antibody are equal. Most antibody is left unbound at these conditions, and this is unlikely to be a useful regime.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	5nM (0.25nM)	10 – 100 nM

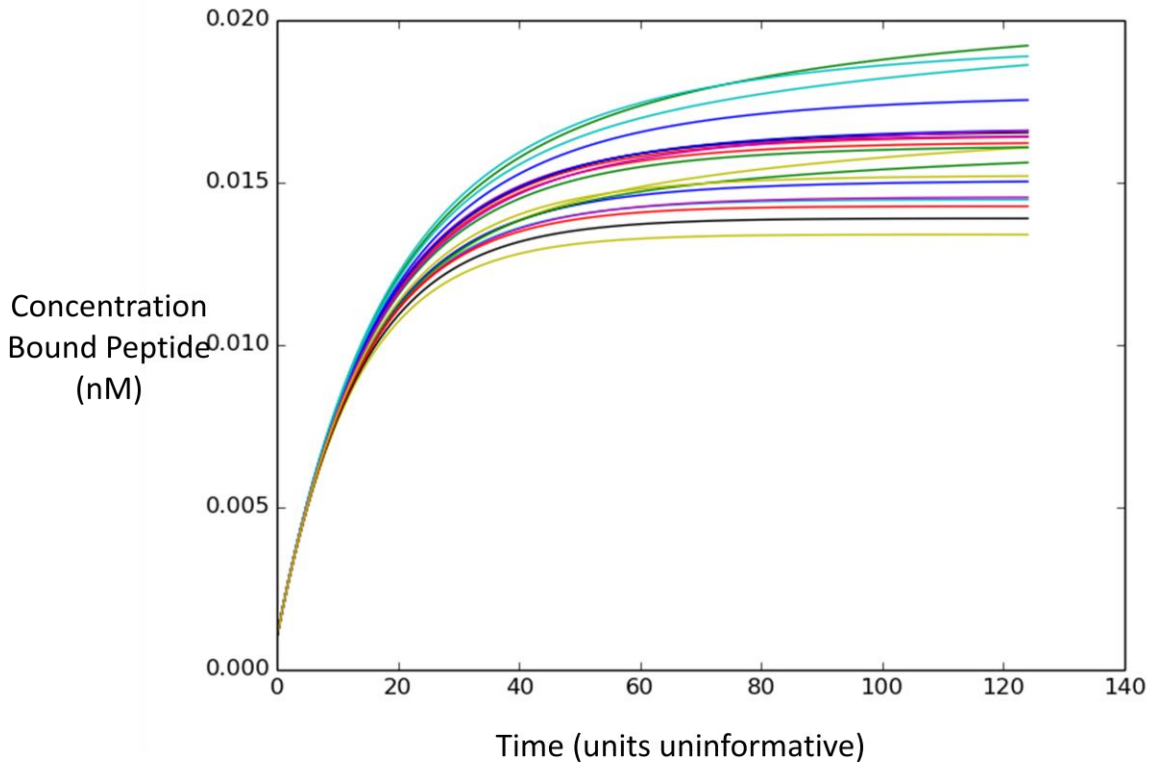


Figure 4.7.5: Hypothesis X , Regime C, Condition 5: Excess antibody, 20 antibodies with 20 peptides

Here peptide concentration remains low, but 100x more antibody are added in order to force binding. As expected, saturation of peptide happens quickly for all affinities.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	500nM (25nM)	5nM (0.25nM)	10 – 100 nM

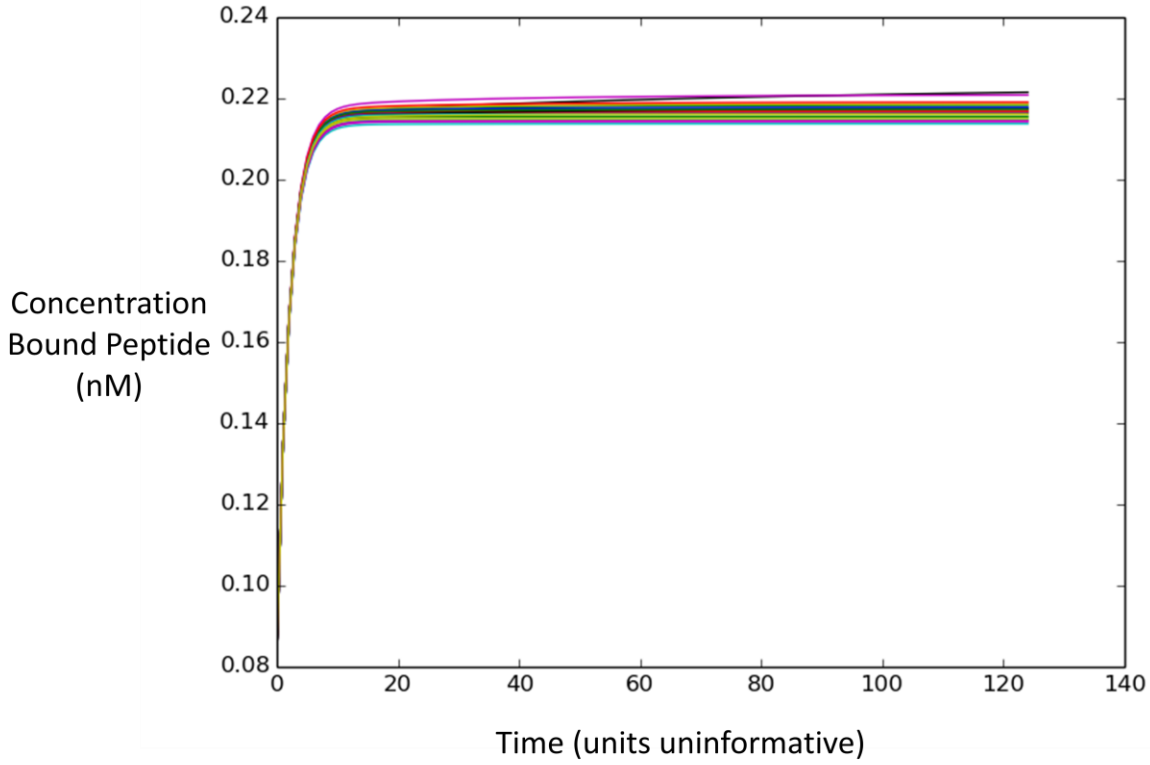


Figure 4.7.6: Hypothesis X , Regime A, Condition 6: 100x excess peptide, 1 antibody with 20 peptides

This test is identical to condition 2, which showed the interesting crossover effect, with the exception that only one antibody is used instead of 20. At this tight range of K_d values, this crossover effect is still observable. As with the 20x20 case, initial depletion time is similar, but there is a longer period of re-arrangements as loadings are focused on the few peptides with the lowest K_d values. This occurs even when the range of K_d values is restricted to the low nanomolar range. This behavior is similar to that observed in **Figure 4.6**, if the 1 hour cutoff was relatively early in the reaction and spot saturation happens at around 0.2 on the graph below.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
1	20	5nM (5nM)	500nM (25nM)	10 – 100 nM

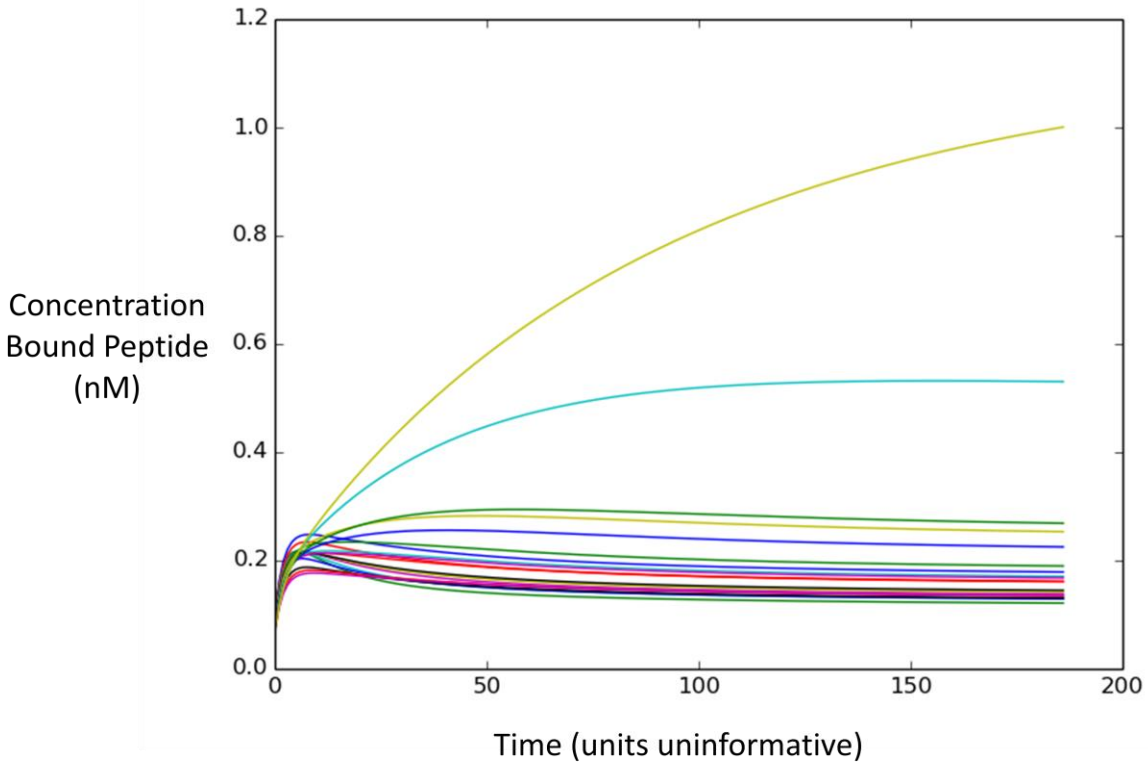


Figure 4.7.7: Hypothesis Y , Regime A, Condition 7: Hyper-Dense Peptide, 20 antibodies with 20 peptides

This situation is extreme, but it illustrates the point that affinity values only inform the meaning of "excess peptide" at which context dependant binding behavior occurs. This looks very much like condition 2, despite the fact that K_d values are an order of magnitude larger.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	50000nM (2500nM)	10 – 100 uM

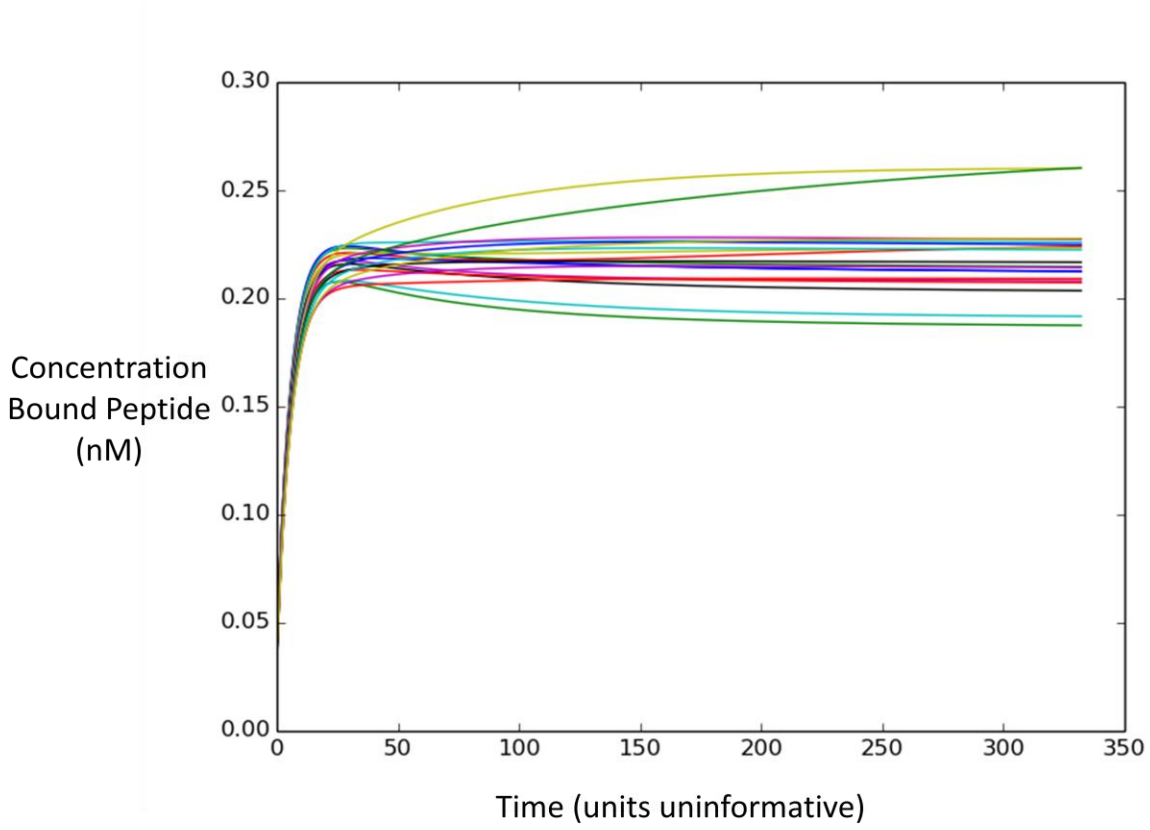


Figure 4.7.8: Hypothesis Y , Regime A, Condition 8: Very Dense Peptide, 20 antibodies with 20 peptides

Here we begin to see a trend. Density of peptide is very important to the behavior of the assay, regardless of what K_d values are. One can achieve resolution of very low affinity interactions if the peptide concentration is high enough, but context dependence becomes less important.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	5000nM (250nM)	10 – 100 uM

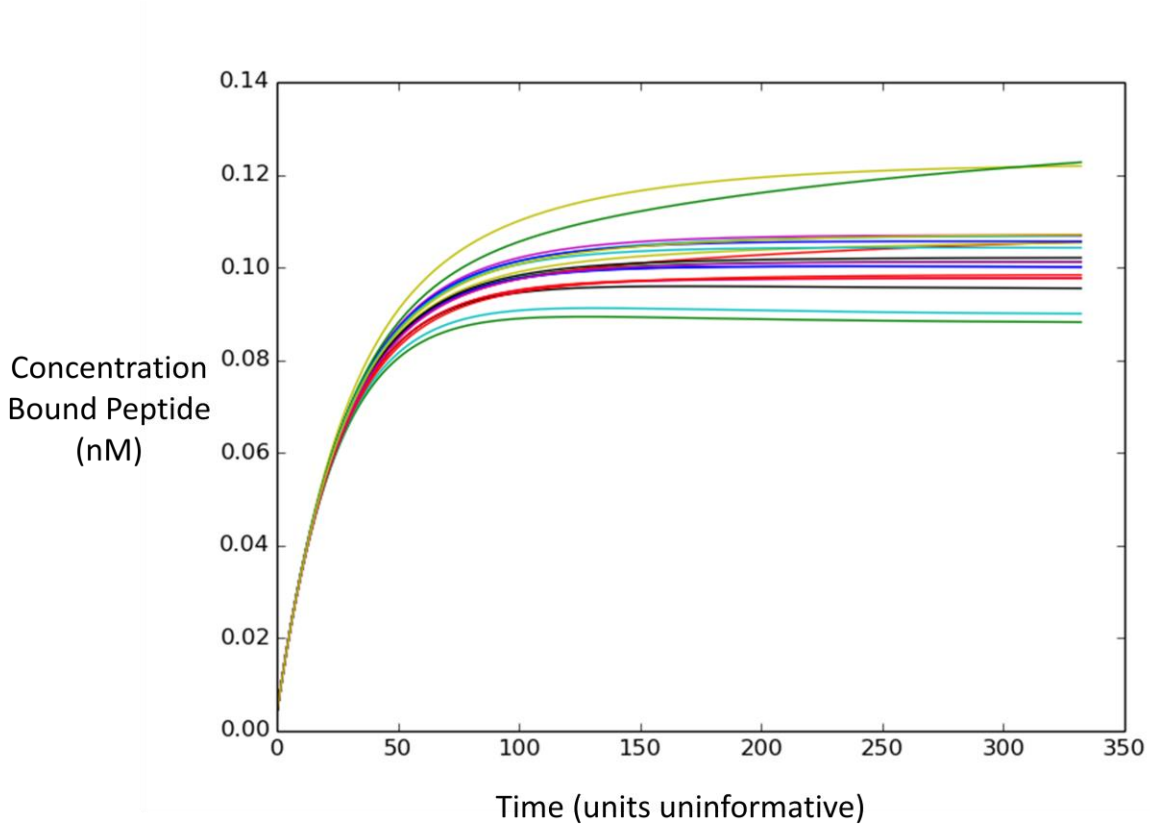


Figure 4.7.9: Hypothesis Y , Regime A, Condition 9: Dense Peptide, 20 antibodies with 20 peptides

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	500nM (25nM)	10 – 100 uM

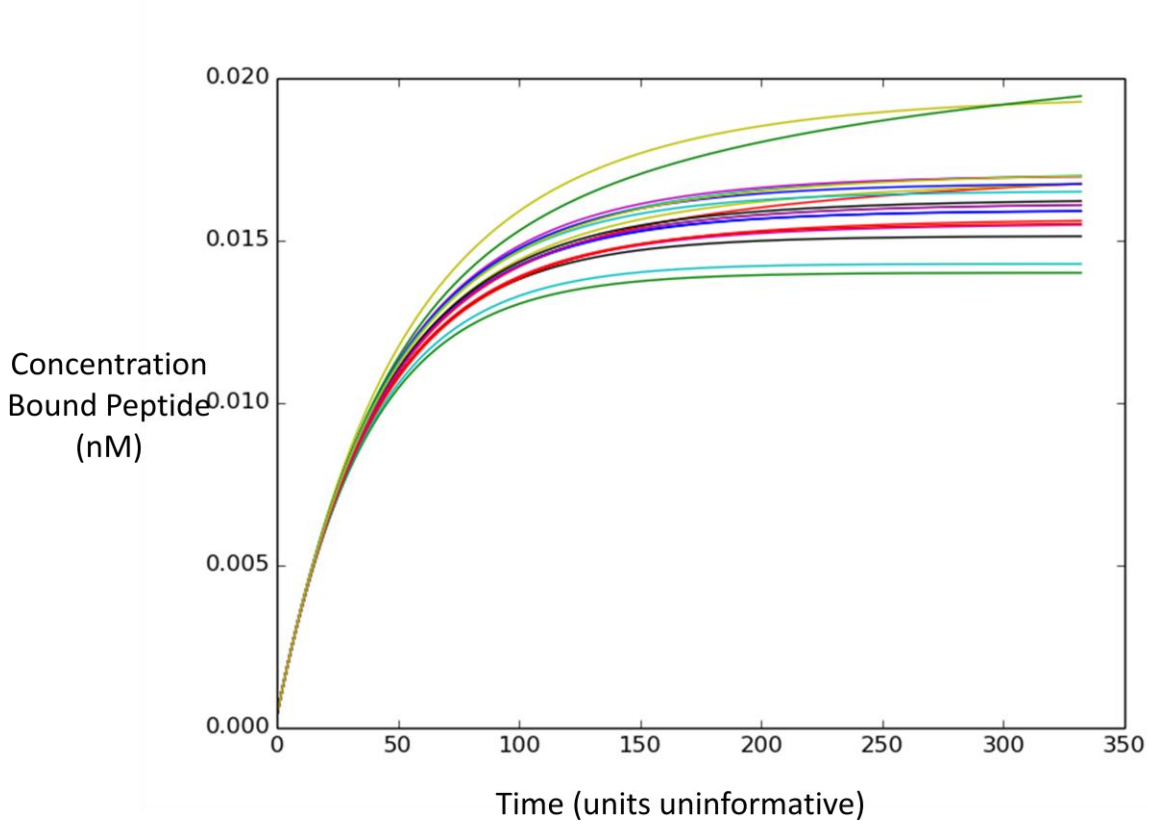


Figure 4.7.10: Hypothesis Y , Regime A, Condition 10: Moderately Dense Peptide, 20 antibodies with 20 peptides

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	50nM (2.5nM)	10 – 100 uM

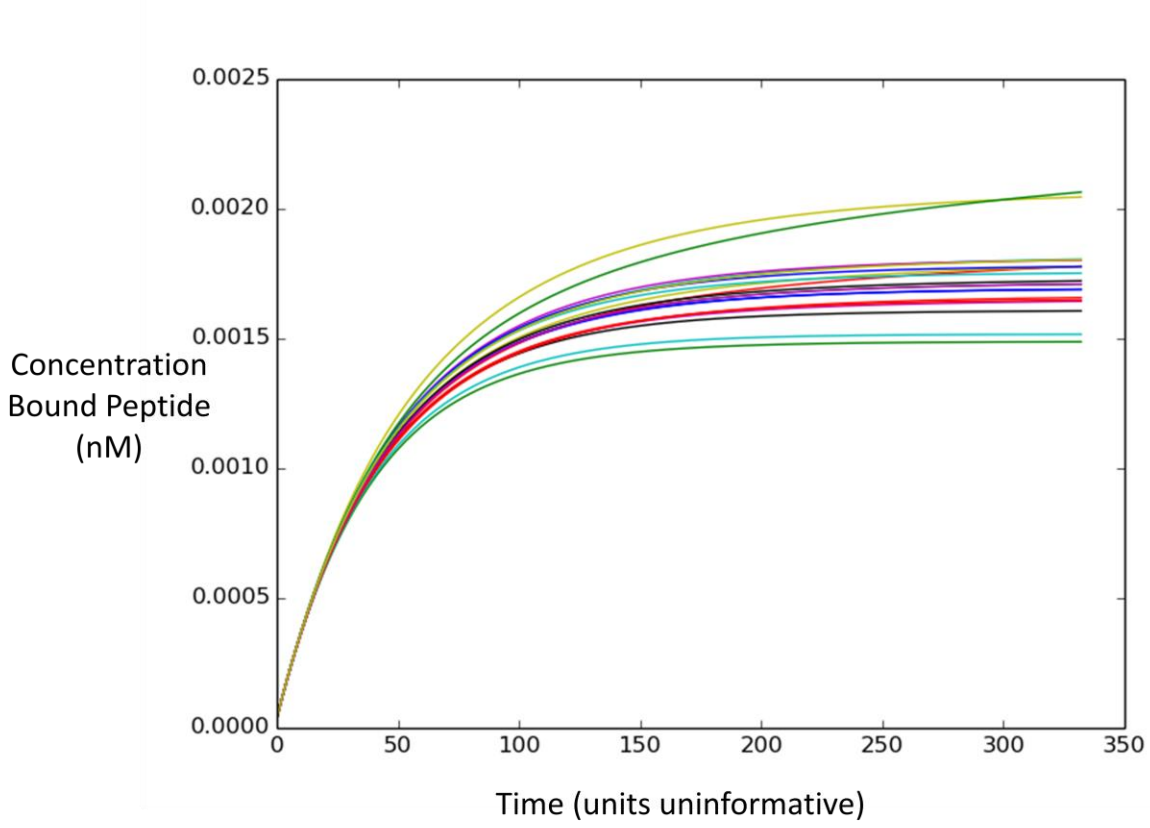


Figure 4.7.11: Hypothesis Y , Regime B, Condition 11: Equality Antibody and Peptide, 20 antibodies with 20 peptides

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	20	5nM (0.25nM)	5nM (0.25nM)	10 – 100 uM

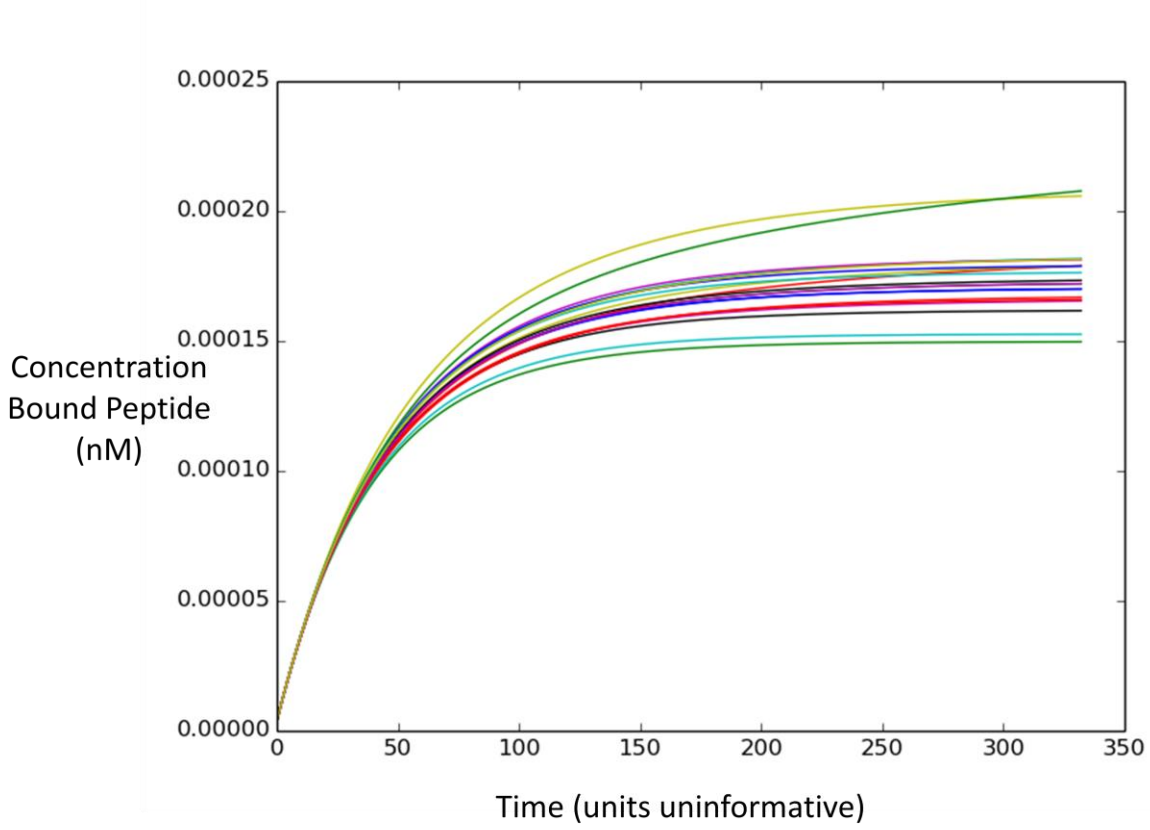


Figure 4.7.12: Hypothesis Y , Regime C, Condition 12: Excess Antibody, 20 antibodies with 20 peptides

N Abs	N Peptides	[Ab_{total}] ([Ab_j])	[P_{total}] ([P_i])	K_d
20	20	500nM (25nM)	5nM (0.25nM)	10 – 100 uM

This is interesting because at low affinities, the assay is much more tolerant of excess antibody, and differentiation can still happen without all peptide sites becoming saturated as was the case in condition 5.

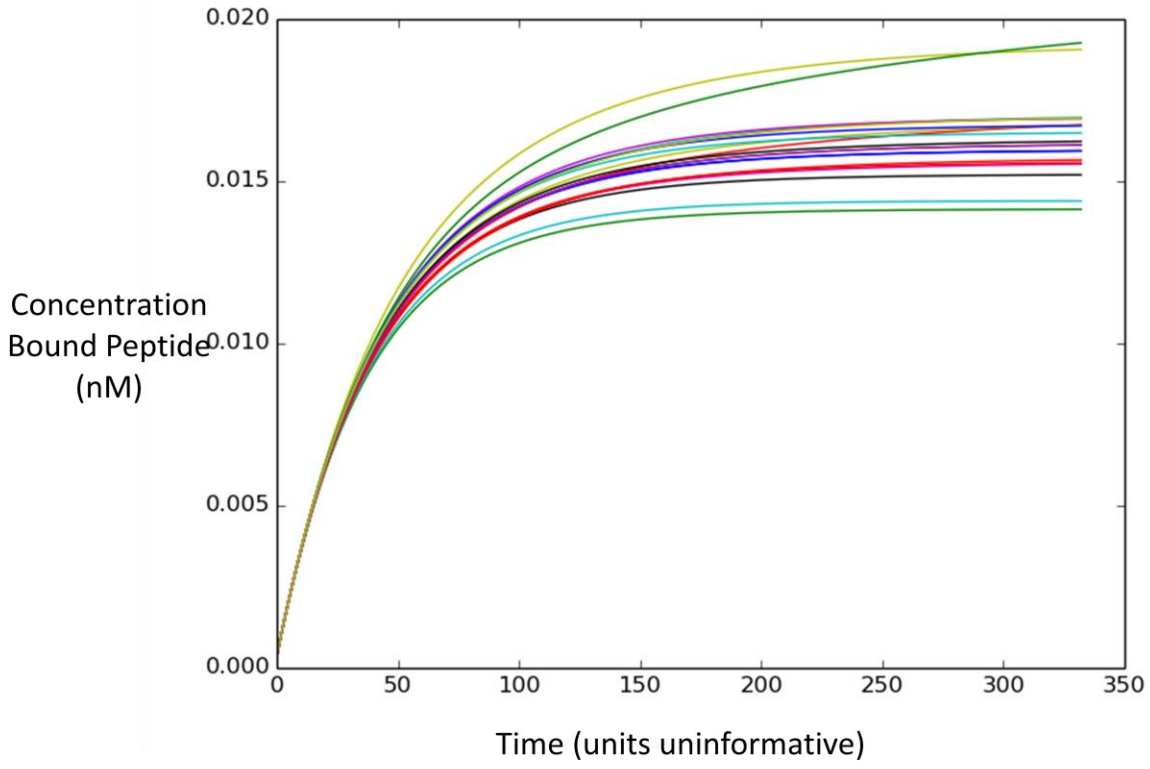
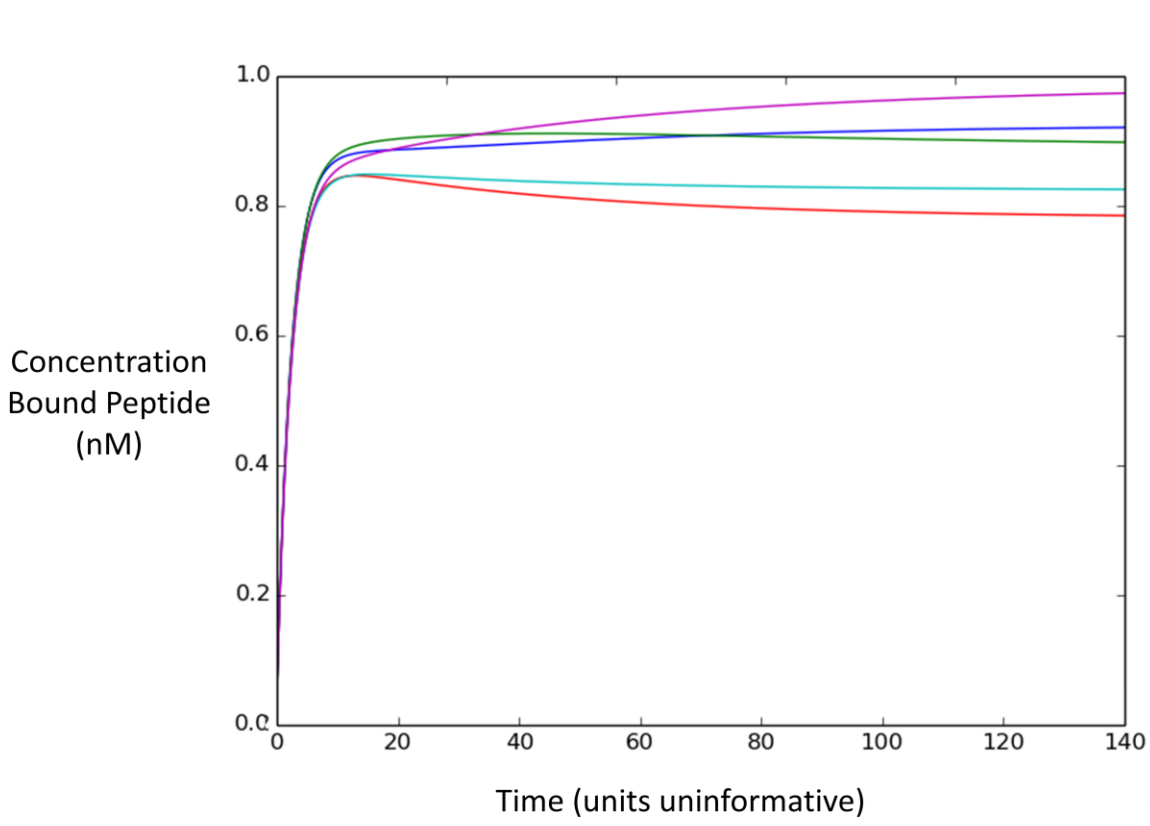


Figure 4.7.13: Hypothesis Z, Regime A, Condition 13: 5 peptides, 20 antibodies

This condition is to test the effect of fewer peptides, but more effective concentration of each while keeping $[P_{total}]$ constant. This is a timely question, as new arrays in our Center are being developed that do just this, and it is an open question how this will effect the ability to do immunosignaturing. Here, there are 5 total peptides each at 100nM, bringing the total to 100nM total peptide (100x compared to total antibody). This seems to bring the assay closer to condition 1, with a quicker saturation time and a wider distribution.

N Abs	N Peptides	$[Ab_{total}]$ ($[Ab_j]$)	$[P_{total}]$ ($[P_i]$)	K_d
20	5	5nM (5nM)	500nM (100nM)	10 – 100 nM



Discussion

This work reviewed existing anomalies in immunosignature datasets, and endeavored to explain these through a kinetic model incorporating multiple antibodies and multiple peptides. A testable prediction related to volume sensitivity was generated under the model, and this was shown to correspond well to experimental data. Then a number of conditions were simulated in order to determine which conditions give rise to context dependant binding and the "crossover effect" which have been observed in many different immunosignature experiments. These conditions also evaluated two competing hypotheses about the K_d ranges likely to be present on the surface of a peptide array.

The kinetic model developed here is very simple and does not account for many aspects of this assay. Diffusion is ignored and peptides are treated as if they are floating around in solution. This seems appropriate for the questions being answered, as the model explains very succinctly some phenomena that were previously puzzling.

First, it is very clear under the model that as peptide concentration decreases, in general less variation in binding is observed on the array. This trend is independent of the unknown K_d matrix in a multi-antibody, multi-peptide experiment. The hypothesis that high peptide densities somehow decrease values in the K_d matrix relative to what they would have been in solution is not needed in order to explain this trend.

Second, this model also explains the context dependant behavior observed from sub-array experiments (Navalkar et al., 2014), superarray experiments (**Figure 4.2**) and in the incubation time experiments (**Figure 4.6)**. Conditions 1, 2, 7, 13, though varying widely in their kinetic and experimental parameters, all showed this effect clearly. This is due to peptide being present in sufficient excess such that antibody is initially depleted.

Once this occurs the peptide-antibody complexes undergo a back reaction and rearrangement. While equilibrium is eventually reached, this is a very different process than that governed by single ligand, single receptor kinetics as given in equation 2. Equation 2 is the intuitive way that most scientists think about binding kinetics, but when multiple receptors and multiple ligands are involved, the laws are instead governed by equations 3, 4, and 5 which result in very different behavior. These previously strange results can be explained by nothing more than simple kinetic theory, and are not surprising.

While the crossover effect and context dependence can be explained by the model, the jury is still out on which hypothesis about the K_d matrix (Hypothesis X or Hypothesis Y) is true. Under Hypothesis X, the crossover effect is observed at a lower peptide density (500nM) than under Hypothesis Y (50,000nM). These numbers should be taken with a grain of salt, since they are dependent on a number of unknown factors, but it is clear that crossover and context dependence would be observable at lower peptide densities if Hypothesis X were true.

Regardless of the content of the K_d matrix, these results imply that a good immunosignature assay depends on antibody-limiting conditions. If antibody is not limiting, peptides will be more likely to saturate at equilibrium (though this depends on the K_d matrix and exactly how limiting the peptides are), resulting in low discrimination ability and low variance. As shown in the volume experiment (**Figure 4.5)** and the corresponding simulations **(Figure 4.4)** variance significantly decreases as peptide concentration decreases relative to antibody concentration. This is a bad thing for immunosignatures, which attempt to resolve minute quantities of antibodies with

potentially low affinities. This could be potentially mitigated by having higher peptide "concentrations" by adding surface sites.

Further experiments might involve a thorough determination of pairwise on and off rates (or at least K_d) for single antibodies with single peptides spotted on a surface. This would determine a detailed K_d matrix. Once this is accomplished for a number of antibodies and peptides, an array could be prepared containing multiple characterized peptides, and incubated with characterized antibodies. Incubating these arrays at various times, volumes and antibody concentrations would validate the model proposed here and inform refinements.

An alternative approach is to skip the pairwise characterization and simply assume a constant K_{on} and a parametric distribution with parameters θ for generating the K_{off} matrix. This matrix would not be measurable directly, so the best that could be accomplished is a maximum likelihood estimate for θ based on observed data. There many methods to fit these parameters to data which are beyond the scope of this investigation.

High throughput assays are important to biology, but the physics underlying them is poorly understood. This paper is an attempt to point others in the right direction for further investigation. Some things are clear: antibody concentration is likely limiting under immunosignature conditions, and these assays likely rely on this fact. Under these conditions, equilibrium complex concentrations depend heavily on context as determined by a K_{on} and K_{off} matrix. The content of these matrices are unknown, and currently the data does not exist to determine these conclusively. This paper simply guessed at these important system parameters, with the goal of pointing out general trends. with proper

experiments, the full behavior of the system could be characterized with the application of designing optimal immunosignature assays.

CHAPTER 5

A COMPOSITIONAL LINEAR MODEL EXPLAINS ARRAY VARIANCE

Abstract

Immunosignatures using non-natural sequence peptide microarrays is a promising technique for disease diagnostics, showing remarkable sensitivity and specificity for several different infectious and chronic diseases including cancer. The mechanism underlying this performance is poorly understood. In order to address this, we fit a simple linear model based on amino acid compositions to several published immunosignature datasets to identify trends, dependencies and enrichments of certain amino acids. Charged amino acids play an important role in the resulting binding profile in all datasets tested, and show a mild ability to predict infections or cancers from normal sera. This cationic charge effect is reduced or eliminated at high pH.

Introduction

Immunosignatures at its core is a technique for quantifying the interactions of a complex analyte (sera containing antibodies with varying specificities) with a chemical space. However, little is known about how antibodies interact with this chemical space and even the very notion of "chemical space" is it this point vague. In the context of peptide arrays, we intuitively understand this as all possible linear peptides, of which a certain subset will bind antibodies. This chapter builds a framework for understanding how antibodies interact with peptides sequences, which amino acids are favored and dis-favored by various monoclonal and serum samples, and how these weightings change across batches (arrays are manufactured in batches).

The antibody repertoire and the process by which it is generated is extremely complex, and encompasses specificities and affinities to a wide variety of antigens (Nobrega et al., 1998). Further, antibodies are known to bind with proteins, sugars, nucleic acids, and small molecules while being negatively selected against self-antigens. Despite this complexity, there may be some sequences or motifs that antibodies generally prefer as an ensemble. So far very little work has been done either in our Center or in the wider scientific community to profile which sequences, amino acid densities and motifs that antibody mixtures prefer.

Grieff, et. al. made an attempt to relate amino acid frequencies to array signals using a very simple linear model on 255 peptides, and found that this explained up to 50% of the signal variance in serum samples, but not in monoclonal samples (Greiff et al., 2012). They predicted an "ensemble effect" whereby antibodies bind peptides based on amino acid content in addition to a specific interaction with a target, this phenomenon being only observed in complex mixtures. With the advent of new arrays, we have the opportunity to test this model on a much larger set of samples and peptides, and assess its usefulness for explaining how antibodies bind peptides, and what factors are contributing to the observed signals.

The method described in (Greiff et al., 2012) is a very coarse grained model. It explains binding in terms of sequence content, but says nothing about sequence order (motifs). Despite the simplicity, it does seem to explain a surprising amount of the variance in array experiments. This warrants a closer look using much larger arrays in both sera and monoclonal samples.

Methods

Enrichment Analysis

Enrichment or over-representation analysis is an approach to finding elements from a discrete distribution that occur more frequently than expected. It is commonly used in microarray, next-gen sequencing and proteomics experiments (Huang, Sherman, & Lempicki, 2009; Subramanian et al., 2005). Typically in high throughput experiments such as these one has a large number of measurements, and a much smaller subset of "interesting" measurements. In the context of peptide arrays, we have a set of peptides S and a much smaller set of interesting peptides S_x determined by their differential or absolute binding. These are the "top n" peptides for a given statistical test. Checking for over-representation of peptides is not very interesting because each peptide on the array is unique. Instead, we compute properties of the peptides which are less unique, such as amino acid or k-mer frequencies. For example, one might select the top 100 peptides on an array by signal intensity, then ask the question "is the subsequence RHSK" appearing in these 100 peptides more than expected by random chance?" This same idea can also be expanded to check all possible subsequences in these top 100 peptides given certain constraints (say all 3 to 5-mers).

P values for each k-mer are calculated according to the hypergeometric distribution, and corrected for multiple hypotheses at 5% FDR.

$$P_i = \frac{\binom{K_i}{k_{xi}}\binom{N-K_i}{n-k_{xi}}}{\binom{N}{n}}$$

P_i	P value for the hypothesis that the ith k-mer occurs in S_x with a frequency predicted by random chance
K_i	Total number of occurrences of the ith k-mer in S
k_{xi}	Number of occurrences of the ith k-mer in S_x
N	Total number of occurrences of all k-mers in S
n	Total number of occurrences of all k-mers in S_x

Table 5.1: Hypergeometric Symbols for Calculating Enrichments

Regression Modeling

A simple linear model predicting intensity (or log-normalized intensity) based on amino acid frequencies is given.

Direct model: $\vec{y} = \boldsymbol{X}\vec{w} + \vec{\varepsilon}$

$\vec{y}$	An $n * 1$ vector of signal intensities, where n is the number of peptides on the array.
$\boldsymbol{X}$	A $n * k$ matrix of amino acid counts, indicating how many of each amino acid is in each peptide, where k is the number of amino acids present on the array plus one dimension for the intercept. Some datasets use peptides which are all the same length. In these cases isolucene is not considered in order to avoid perfect multicollinearity of variables.
$\vec{w}$	A vector of weights, fit by ordinary least squares.
$\vec{\varepsilon}$	An error vector, assumed to be normally distributed.

Table 5.2: Explanation of variables used in linear regression model

This amino acid linear model (AALM) is an attempt to explain array binding only through amino acid frequencies. When fitting the above model to the CIM10K data, the equation must run through the origin. This is because each peptide on these arrays is the same lengths, resulting in perfect multicollinearity and a useless model if an intercept is used.

Monoclonal Spiking Experiment

Normal mouse serum was diluted 1:500. This solution was aliquotted into six vials. To five of the vials, differing amounts of affinity purified anti-GFOD polyclonal antibody was added such that vials contained 0.1nM, 1nM, 2.5nM, 5nM and 10nM of the

spiked antibody in addition to the diluted serum. The sixth vial contained only serum as a negative control, and a seventh vial contained only 40nM anti-GFOD as a positive control. Each of these seven prepared sample was incubated in duplicate on the CIM10K-V2 array for 1 hour primary incubation and 1 hour with anti-mouse IgG conjugated with Alexa-647 dye as a secondary and washed. Arrays were scanned with an Agilent-C microarray scanner and aligned with GenePix Pro 6.0.

Datasets

Several datasets were compiled to assess the linear model and enrichment strategy across several batches and experiments. These include:

Chip Set and Samples	Description
HT152 – Sera IgG	seven infectious disease cohorts and normal sera. Over 330,000 peptides synthesized using NSB-9 wafers (3nm distance between peptides within a spot)
HT22 – Monoclonal IgG	eight monoclonal antibodies. Over 330,000 peptides. These peptides are longer than HT152 peptides and synthesized on standard wafers.
HT22 – Sera IgG	seven infectious disease cohorts and normal sera. Over 330,000 peptides. These peptides are longer than HT152 peptides.
CIM7-18 – Sera and Monoclonal IgG	small cohorts including monoclonals, normal, and three infectious diseases. About 10,000 peptides.
CIM10K – Dengue v Normal v WNV v Malaria	Four small cohorts run on CIM10K arrays, Spring 2013. About 10,000 peptides.
CIM10Kv3 – Dog Cancers	Cohorts of dogs with cancer (used for AA enrichments only). About 10,000 peptides with some D-amino acids mixed in

Table 5.3: Description of peptide array datasets used in this study: HT152 and HT22 and CIM7-18 each contain over 330,000 random sequence peptides, but each contains a different library of peptides. CIM10K and CIM10Kv3 both have over 10,000 random peptides, but CIM10Kv3 has dextro amino acids.

Results

There are two analyses that were performed on several immunosignature datasets. The first involved fitting a linear model (AALM) to each dataset and testing various hypothesis about the model. The second involved calculating sequence enrichments for several lists of selected peptides. These methods are related in that they both explore the role of sequences in determining observed binding.

Compositional Linear Model Fitting

This assay shows binding of serum antibodies to a large number (tens to hundreds of thousands) of addressable peptides. An obvious question is, which peptide sequences, amino acids, and motifs do antibodies prefer to bind in general? Does this vary between batches, samples, cohorts and protocols? Almost nothing is known about which sequences monoclonal antibodies and sera prefer.

The first question is whether a simple compositional linear model can explain array binding at all. **Figure 5.1** shows compositional model (AALM) fit quality to many samples on several array batches. Serum samples have R^2 values ranging from 0.3 to 0.7, while monoclonal antibodies tend not to fit well ($R^2 < 0.1$). There were no significant differences in model fit between classes (e.g. Dengue vs. Normal), or for IgM vs. IgG, but there were significant differences between array batches.

Coefficient values for each amino acid varied both between and within batches. Within batches coefficient variation was dominated by the cationic effect (response to positively and negatively charged amino acids) (**Figure 5.2)**.

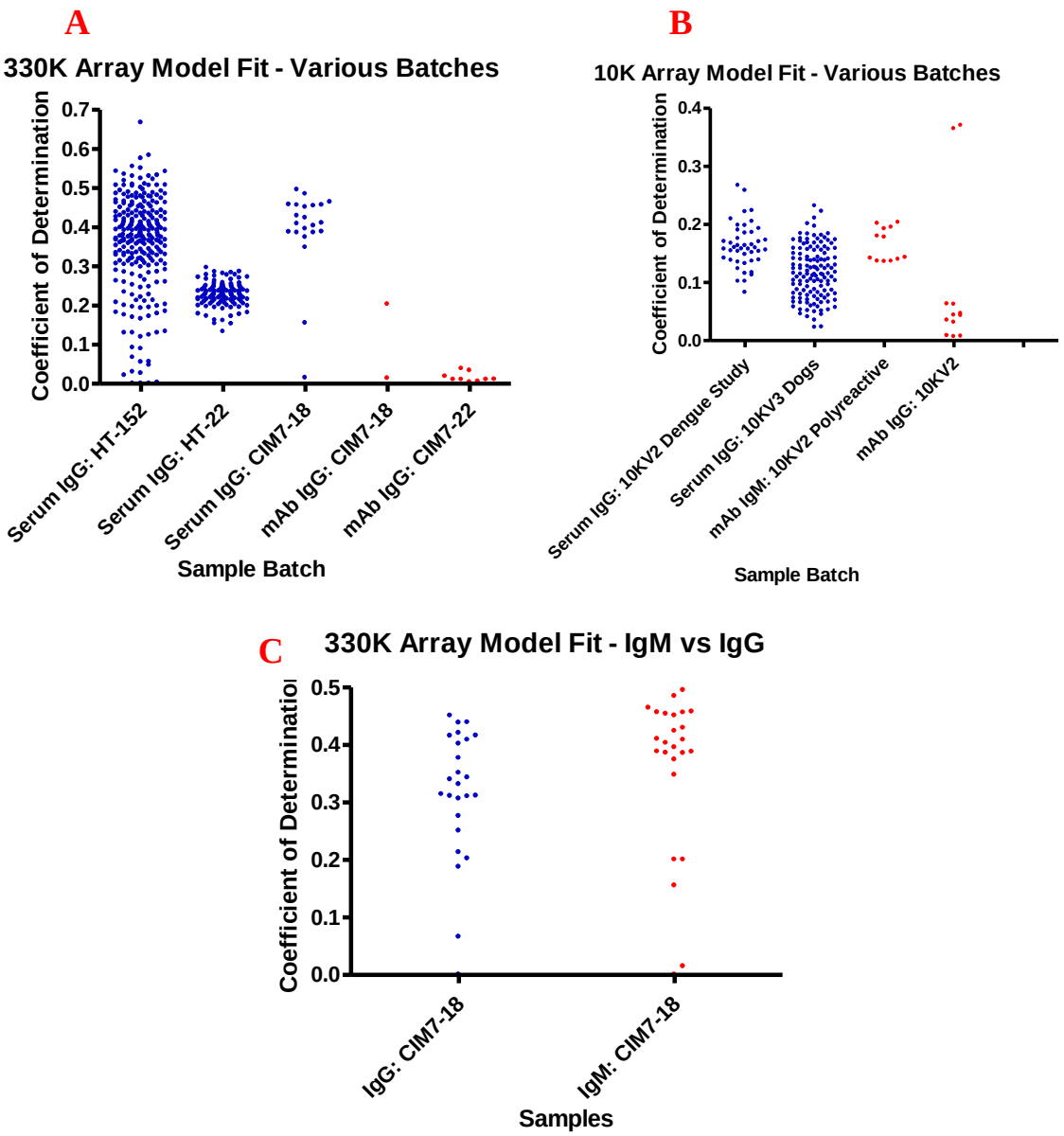
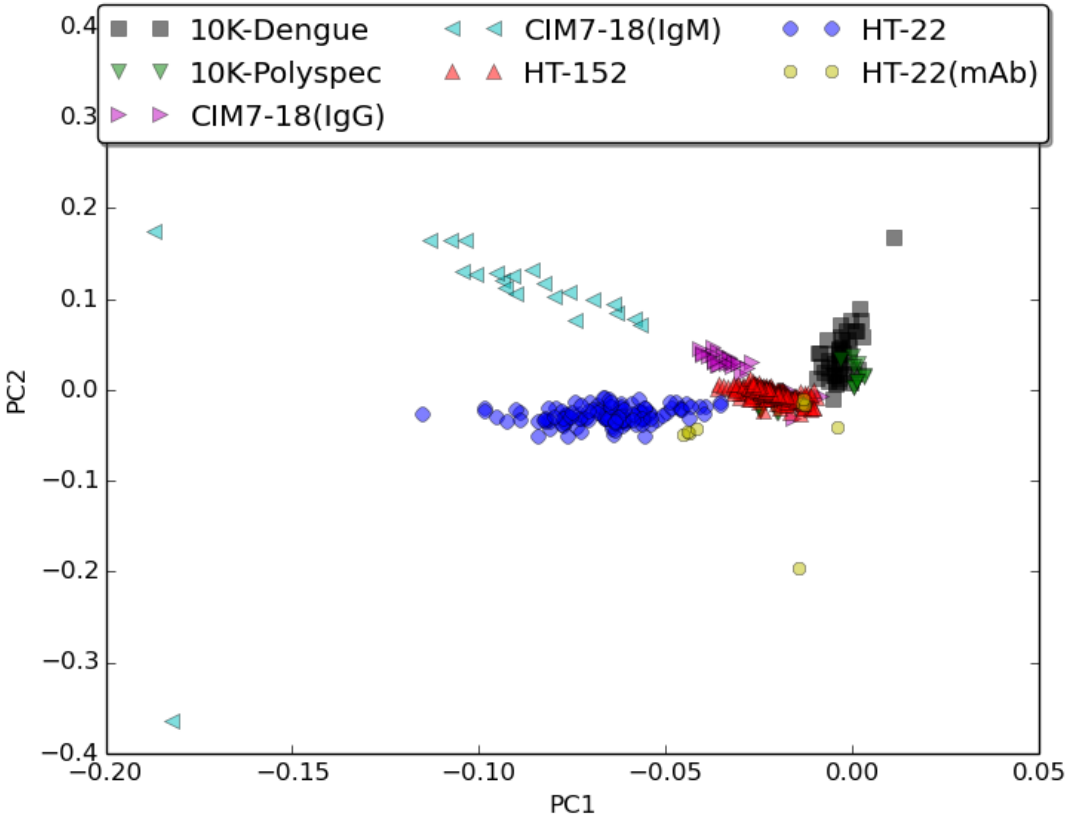


Figure 5.1 - Coefficients of determination for AALM: These plots show R^2 values for AALM on different batches and biological samples. These range from around 0.2 to 0.6 for serum samples, and do not fit well with monoclonals ($R^2 < 0.1$). Polyreactive monoclonal antibodies fit the AALM better than high affinity monoclonals, and 330K arrays seem to drive a better fit than 10K.

5.2A


10K-Dengue
10K-Polyspec
CIM7-18(IgG)
CIM7-18(IgM)
HT-152
HT-22
HT-22(mAb)
PC1
PC2
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.20
−0.15
−0.10
−0.05
0.00
0.05

5.2B

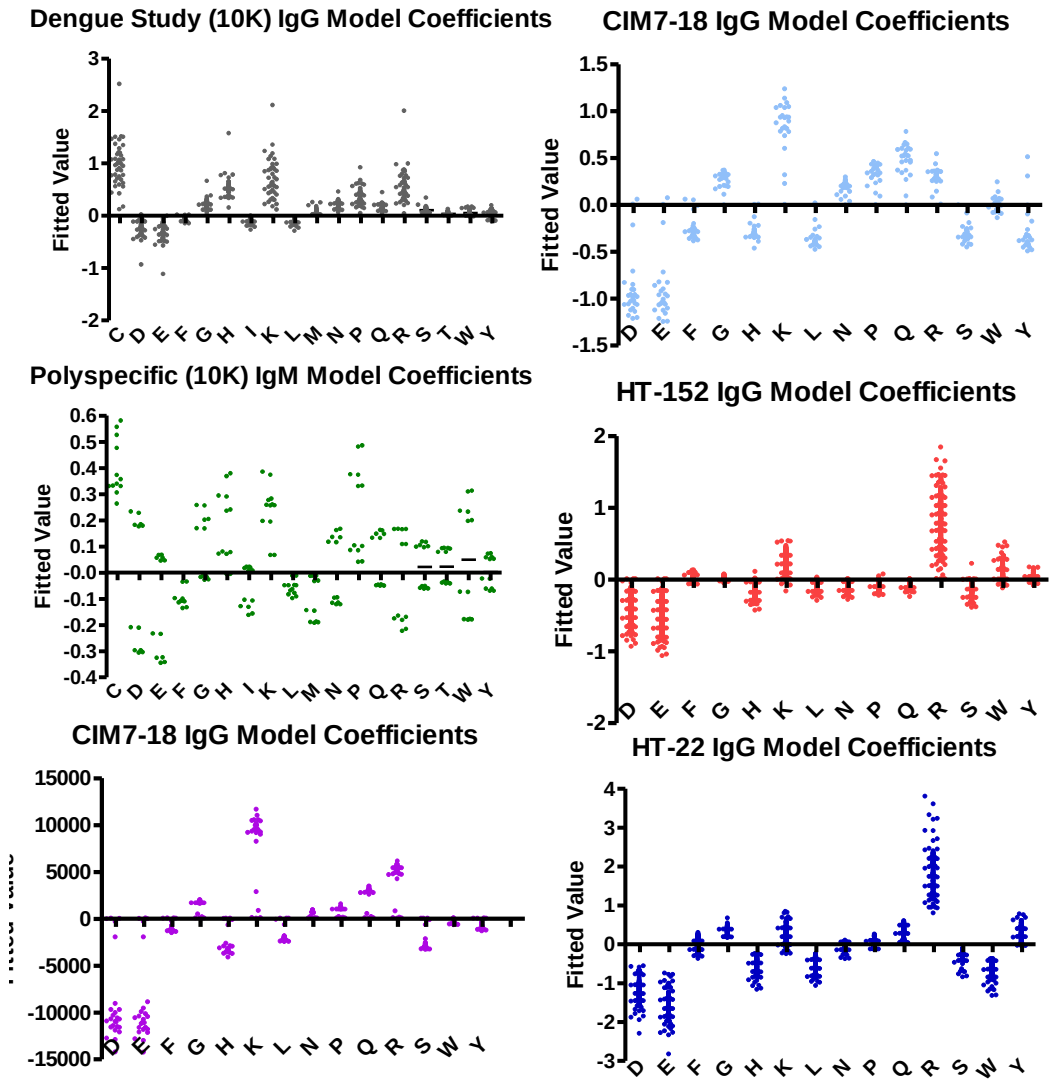


Figure 5.2: PCA and Plots of AALM Coefficients: (5.2A) principal component plot reveals batch specific effects on several amino acid axes, most notably the R/K and E/D axis. This is particularly apparent in the HT-152 and HT-22 datasets, where samples are splayed out in a line reflecting variation in response to charged residues These wafers also worked best so far as a diagnostic in infectious diseases (data not shown) . (5.2B) These are model coefficients for each dataset tested, and show a view of the data without dimensionality reduction.

The importance of charged amino acids in both between batch and within batch variation begs the question regarding the effect of pH on the amino acid binding profile. An experiment was designed to test the pH effect on residue specific binding, where a single serum sample was run at varying pHs from 6 to 9. Higher pH values resulted in lower binding overall for both IgG and IgM. This effect was most heavily loaded on the charged amino acids, particularly Arginine and lysine (**Figure 5.3 top**).

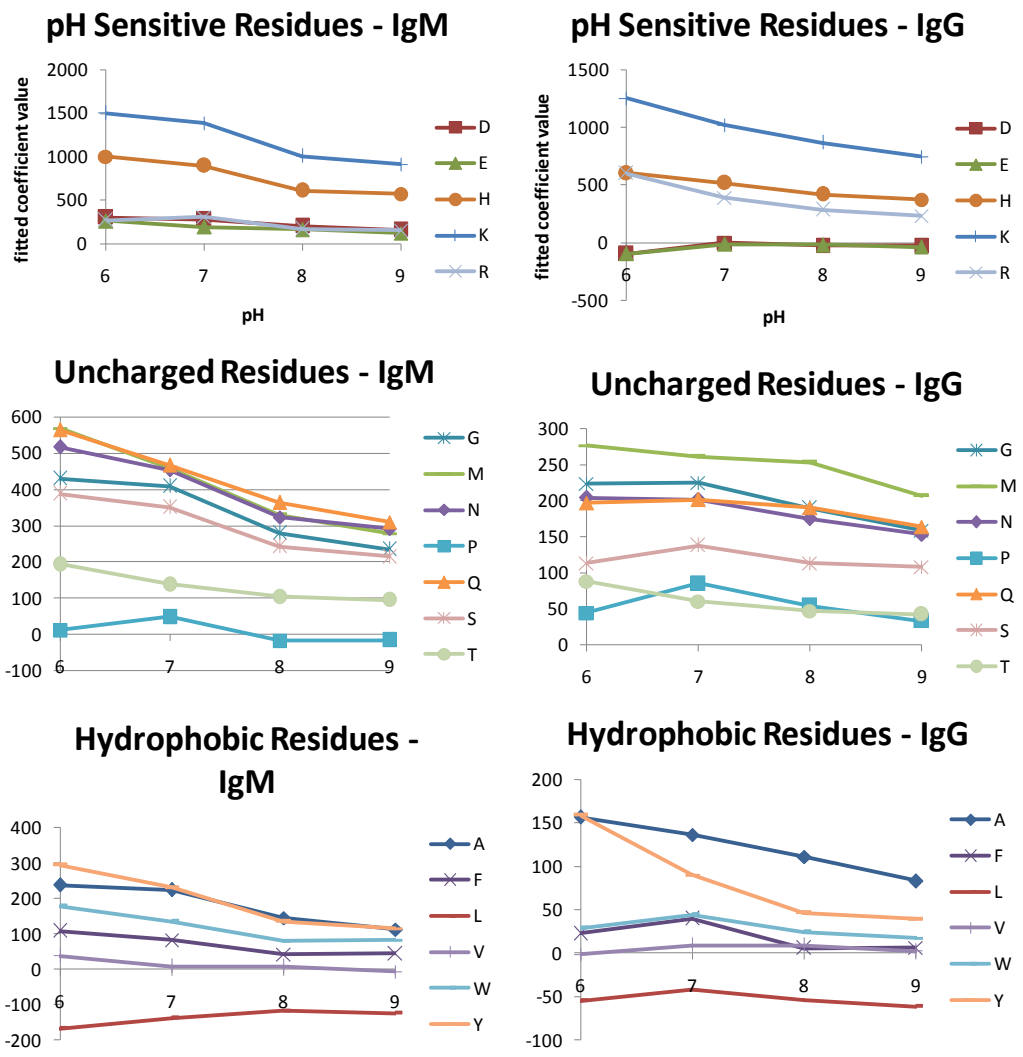


Figure 5.3: AALM Coefficients at Varying pH in Normal Sera: These plots show the coefficient trends as pH changes. There is a general downward trend in binding overall, most significantly involving the anionic residues aspartic and Glutamic acid.

Perhaps a more direct measure of the pH sensitivity involves the isoelectric point of the peptide. Here, instead of quantifying the loadings on each residue, pI is simply plotted against pH to reveal any dependence. Overall binding did not change significantly between pH 9 and pH 6 for IgM, but IgG showed significant changes especially among

the peptides with higher pI (**Figure 5.4**). The AALM and pI results together suggests pH has an important effect on IgG binding to the basic amino acids lysine and Arginine.

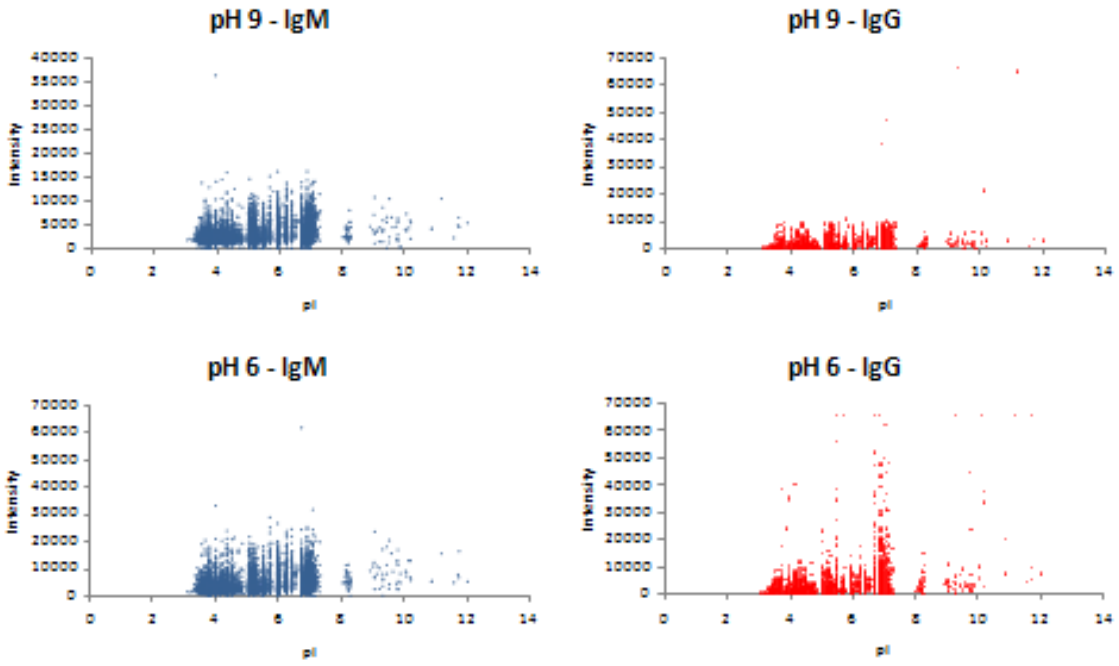


Figure 5.4: Effect of Isoelectric Point on Binding Intensity to Normal Sera at Varying pH: Plots showing pI (x axis) versus intensity for a normal serum sample at pH 6 and pH 9. IgG shows significant pH sensitivity, with higher reactivity to peptides with pI greater than 6.

Compositional models like AALM are also useful for revealing quality issues with arrays. Wafer HT-22 had issues with anomalous histidine binding, which becomes apparent when coefficient loadings are examined (**Figure 5.5**). Monoclonal antibodies which should not fit the model showed consistant loadings on histidine. Surprisingly, this did not seem to affect the ability to determine epitopes, as several of the histidine binders

(DM1A, AB1) also showed very clean epitopes when motif detection was applied to these data.

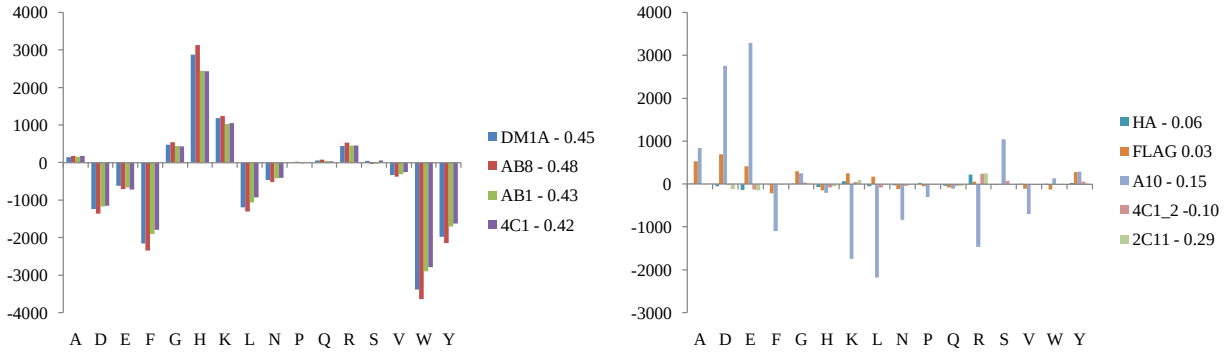


Figure 5.5: Normalized linear model coefficients from wafer HT4-22 for all nine tested monoclonal antibodies: Four of these (left) were strong N-terminal histidine binders which resulted in a high R^2 value. This seems to be driving the fit, and was a known problem with this batch of arrays. The other five monoclonal antibodies had a very poor fit, with R^2 values in the range of ~0.02 for the direct AALM.

These results are in agreement with those previously reported by (Greiff et al., 2012). Complex mixtures of antibodies fit a linear model fairly well, while monoclonal antibodies do not. This preference for sera warrants further investigation, particularly to answer the question of whether there is any disease specific effect on coefficient loadings. To test the effect of a disease perturbation of the antibody mixture, a spiking test was devised intended to mimic the effects of an infection causing production of immunodominant antibodies (Frank, 2002). 1:500 diluted healthy BALB/C mouse serum was spiked with differing amounts of affinity purified anti-GFOD antibody. It can be seen from this experiment that R^2 decreases dramatically as more anti-GFOD is spiked in (**Figure 5.6)**.

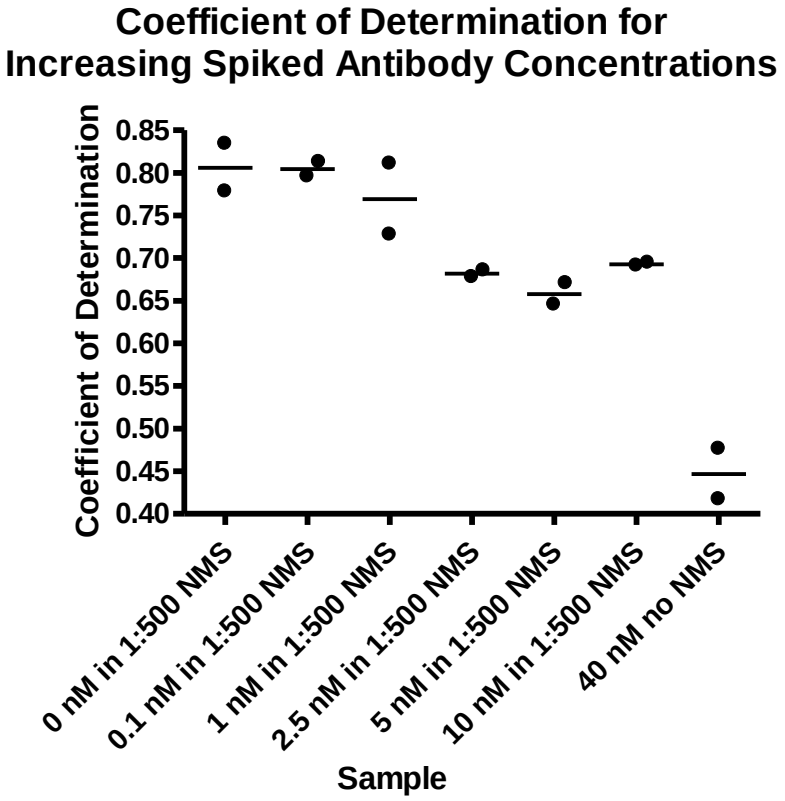


Figure 5.6: Coefficients fitted values GFOD Spiking Experiment: 1:500 diluted normal mouse serum (NMS) was spiked with varying quantities of anti-GFOD affinity-purified polyclonal antibody. Increasing concentrations of spiked antibody dramatically reduce fit, even when the same concentration of NMS is still present in the background. Biased solutions of antibodies have a fundamental difference in binding as compared to unbiased solutions, and this may be relevant in disease diagnostics.

Another interesting question is whether loadings change significantly between batches or array samples. The same normal human serum sample (ND134) was run on multiple array batches for quality control. When these data are fitted independently to the AALM, it is clear that loadings are dependent on the production batch, though there is also a clear and consistant trend **(Figure 5.7)** with respect to these loadings. The effect of basic residues Arginine and lysine is again clear, but surprisingly this effect is at times reversed. In batch 118, lysine and Arginine have a negative contribution to binding, while

Glutamic and aspartic acid exhibit a positive effect; the opposite of what is normally observed.

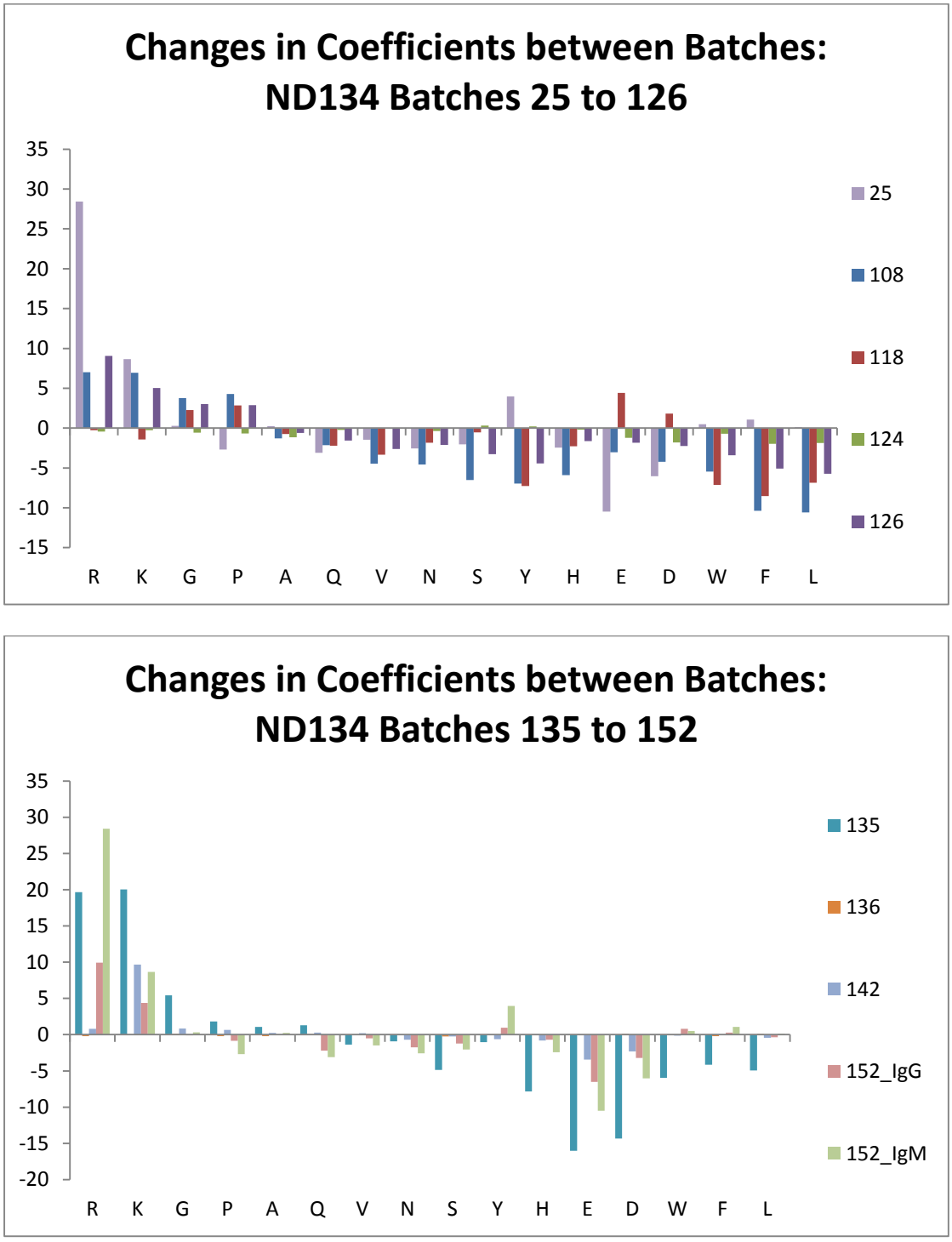


Figure 5.7: Coefficients fitted values across batches for Normal Donor 134: One normal sample was run across multiple batches of the array and fit to the log-transformed AALM, resulting in a general trend of amino acid binding. R^2 values ranged from 0.72 to

0.93. A charge reversal effect is seen in batch 118 involving R,K,E and D, indicating that assay conditions (probably pH) have a strong effect on the amino acid binding profile.

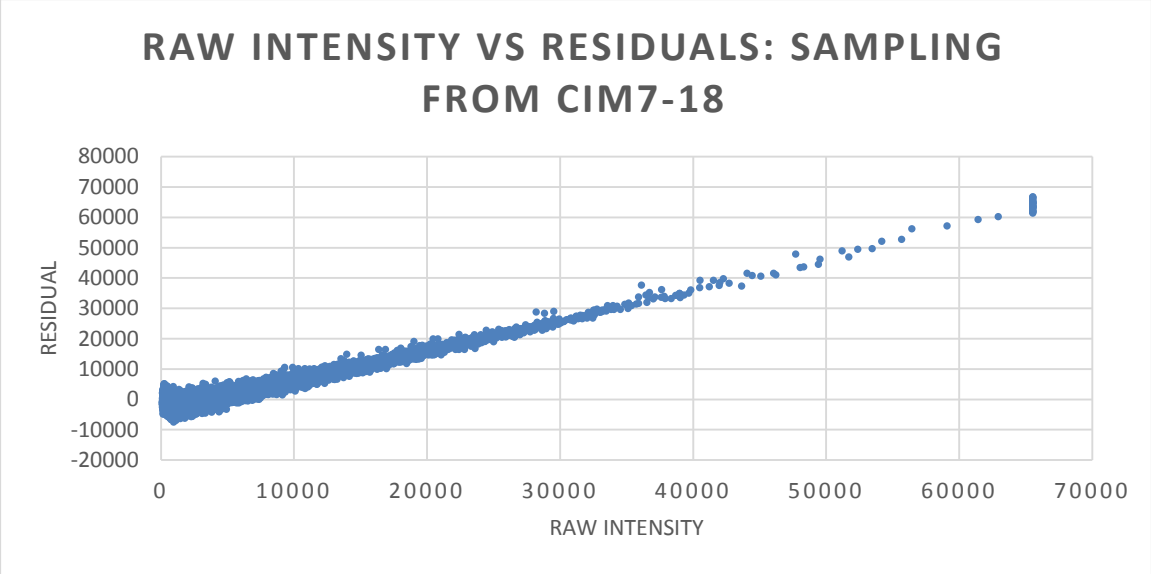


Figure 5.8: Residuals for AALM fit on CIM7-18 Wafer: Residuals reflect a linear trend, indicating that a linear model does not explain the variance equally over the range of intensity values.

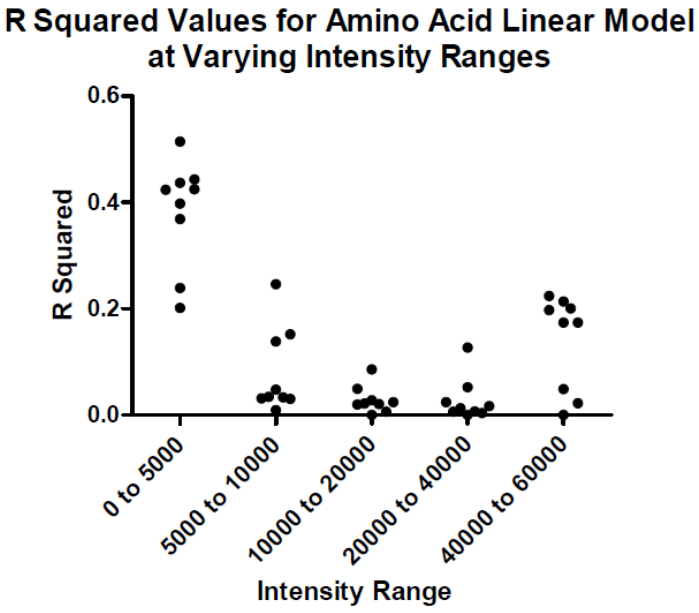


Figure 5.9: Coefficients of Determination for Piecewise Discontinuous Model: Model fit is not distributed equally amongst the intensity range. Low and high intensity peptides

fit better, and while there are fewer peptides in the middle range, there are over 10,000 peptides in each tranche.

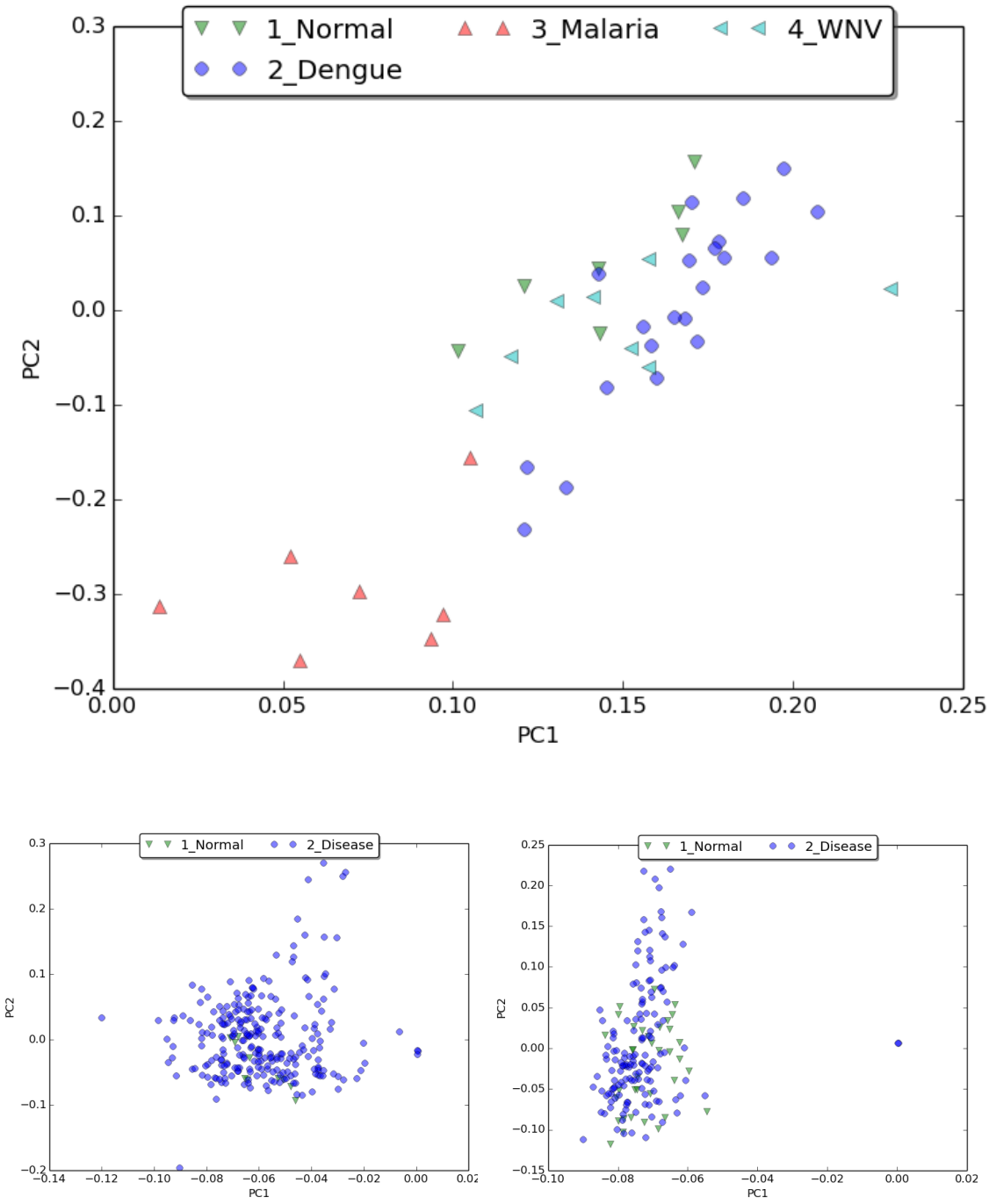


Figure 5.10: Principal Component Analysis on Model Coefficients – Three Datasets: Log-transformed AALMs were fit to each sample from CIM10K (top), HT152 (bottom left) and HT22 (bottom right). The two bottom studies involved seven infectious disease cohorts and one normal cohort, while the top study was done on smaller arrays with three disease and one normal cohort. In all three cases, separation can be seen between normals

and some diseases, though in the two 330K studies this effect is subtle. The effect is pronounced in the 10K study, where malaria samples clearly separate from the other cohorts. This might be evidence for cohort specific variation in terms of the AALM, or it could be due to systematic sample effects due to the fact that each cohort came from a different supplier.

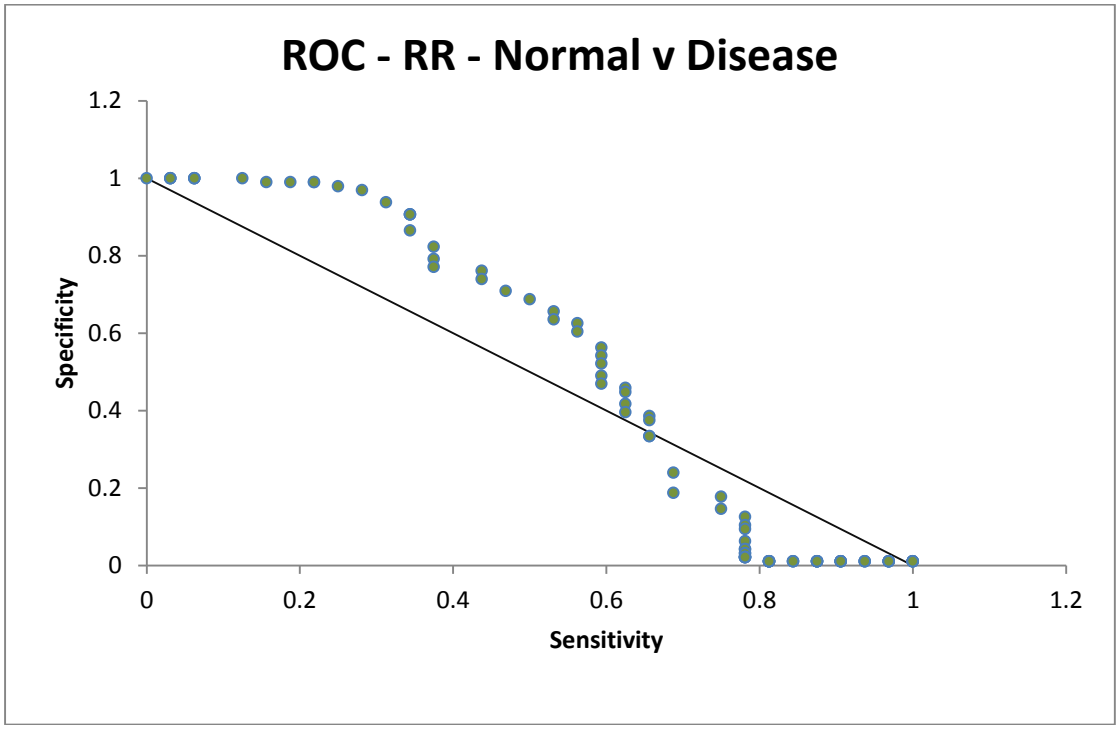


Figure 5.11: ROC Performance (Normal vs. Disease) on Wafer HT-22 for Dipeptide Motif RR: ROC analysis on the dipeptide motif RR reveals a weak predictive ability for normal samples, indicating the possibility of an ensemble difference between normal and infected samples in terms of their response to Arginine residues.

Enrichment Analysis

The AALM fit in the previous section reveals a strong role for charged residues in binding determination. It would be interesting to know whether these residues (or others) are enriched in selected peptides used for immunosignature diagnosis. Very little is known about the sequences selected for prediction, as no comprehensive statistical analysis of these sequences has ever been reported. Enrichment analysis was conducted for three disparate immunosignature datasets. These were intentionally selected to cover a wide variety of platforms and samples to see if any general trends arose. Charged amino acids were enriched in all three tested datasets, but cationic ones were only enriched in the two human datasets, while dogs showed enriched anionic residues. Surprisingly, proline was enriched in all three datasets, though only slightly in dogs. These results are summarized in **Figure 5.12**. Further analysis revealed these enriched charged amino acids exist only in the down peptides (lower in disease than in normal). This is a remarkably stable phenomenon, present in all three diverse datasets (**Table 5.4**).

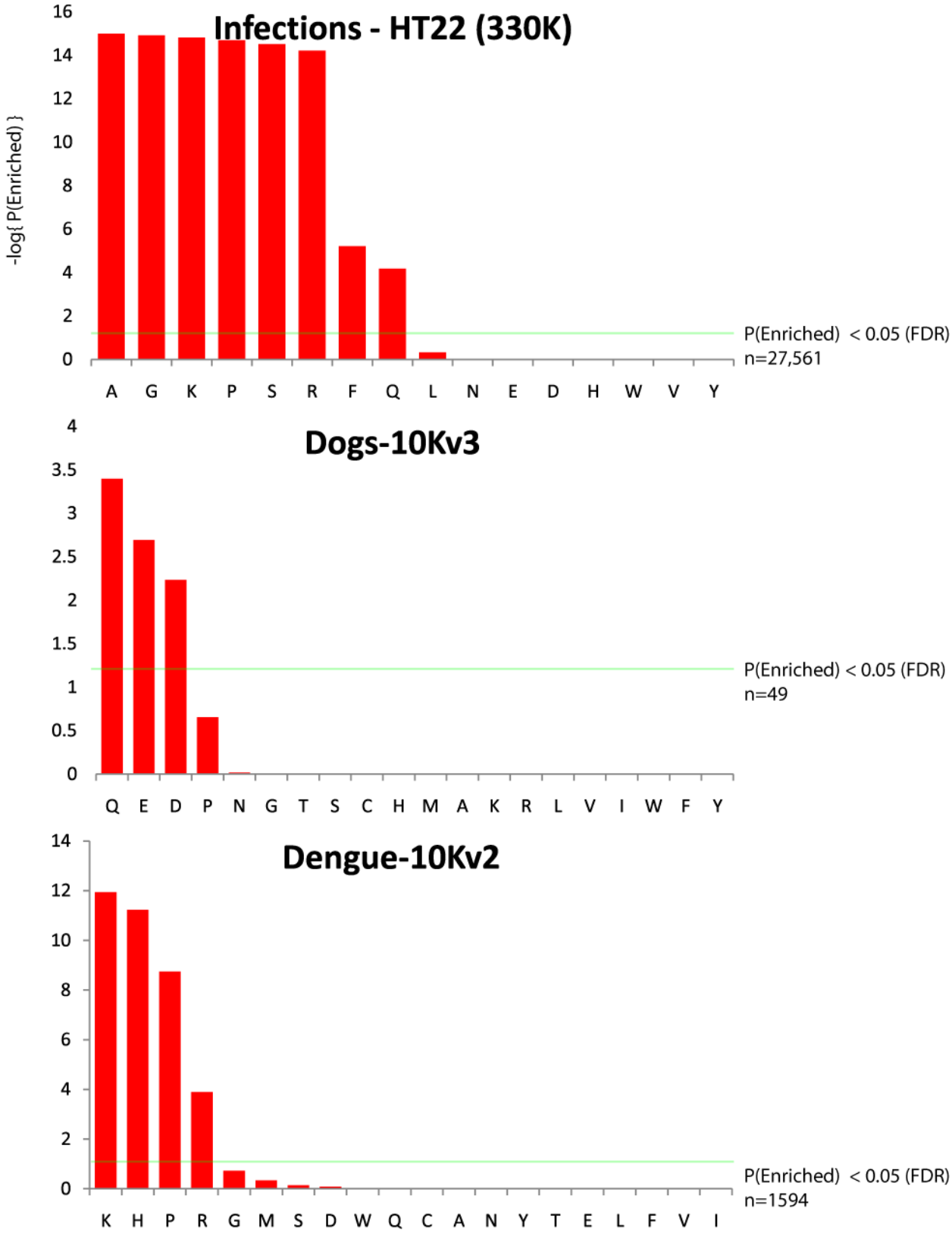

Infections - HT22 (330K)
-log{ P(Enriched) }
16
14
12
10
8
6
4
2
0
A G K P S R F Q L N E D H W V Y
P(Enriched) < 0.05 (FDR)
n=27,561
Dogs-10Kv3
4
3.5
3
2.5
2
1.5
1
0.5
0
Q E D P N G T S C H M A K R L V I W F Y
P(Enriched) < 0.05 (FDR)
n=49
Dengue-10Kv2
14
12
10
8
6
4
2
0
K H P R G M S D W Q C A N Y T E L F V I
P(Enriched) < 0.05 (FDR)
n=1594

Figure 5.12: Enrichment P Values for Each Residue, Three Experiments: Three immunosignature datasets (all IgG based) were tested for enrichments within significant peptides (Normal vs. Disease, FDR < 0.05). Cationic amino acids were highly enriched in the two human datasets (Dengue and Infections), while anionic ones were enriched in dogs. Proline was significantly enriched in the human datasets, and nonsignificantly (slightly) enriched in dogs. In general, charged amino acids and prolines are often enriched in immunosignatures. These results agree with and confirm the principal component plots in **Figure 5.10**, which illustrate this phenomenon in terms of AALM coefficients.

Discussion

A variety of datasets encompassing monoclonal and immunosignature experiments in humans and dogs were compiled in order to do a meta-analysis of sequence dependencies with the goal of elucidating the role of sequence composition in immunosignatures. First, a compositional amino acid linear model (AALM) was constructed and fit to each dataset, and loadings for each amino acid were examined and compared across datasets. Then, enrichment analysis was conducted on the same datasets in order to determine which amino acids appeared more often than expected in immunosignatures. These findings indicated a strong role for charged amino acids and possibly pH. Thus, an experiment was designed to test the effect of pH on binding profile.

The results from this meta-analysis indicate a strong role for charged residues in determining final binding intensity. This dependency is most clearly visible on the low end of the intensity spectrum (0 to 10000) and also at the very high end (~65000). In both these ranges, a simple linear model explains up to 50% of the within-array variance (**Figure 5.9**). Considering the high noise in these datasets (correlation between technical replicates averages 0.85), this is a remarkably good fit.

The errors are clearly not normally distributed (**Figure 5.8**), indicating there are additional unmodeled intensity-dependant variables contributing to binding. This indicates there is an opportunity to explain more variance in terms of additional latent variables.

While it is clear that charged amino acids are important determinants of array binding, it is less clear whether these models can inform and improve immunosignatures. It is unlikely that compositional models such as the AALM can inform disease vs. disease classification (**Figure 5.10**), there is some evidence that it could predict normal vs. disease **(Figure 5.11)**. This result is rather surprising, as one would expect that binding based solely on charge differential would be rather non-specific. There have been several previous studies in our lab (Muskan Kukreja, Kurt Whittemore, unpublished data) showing a mild predictive ability for disease vs. normal, and we have termed this strange phenomenon the AbStat, or immune temperature. The results shown here indicate that the AbStat may be due in part to a difference in charge affinity between normal and disease sera. This is very surprising and unintuitive, because if it is true, it implies we are measuring a previously undiscovered kind of network effect on the humoral immune system. Perhaps a nonspecific class of antibodies exists with a similar role as antimicrobial peptides. It is also interesting that the amino acids that show up enriched consistently in the down peptides are the same ones which appear en-masse in antimicrobial peptides (K,R,P,G) (Zasloff, 2002).

These charged amino acids are strongly affected by pH in IgG based assays. As pH increased, cationic amino acids contributed less to binding (**Figure 5.3**) and anionic amino acids were largely unaffected. Further, there was a pI specific drop in intensity at

high pH assays **(Figure 5.4)**. This is expected, and suggests a way to reduce or eliminate the cationic effect in these assays. It is unknown how pH affects immunosignatures.

The cationic effect observed on these arrays has interesting properties, and the presence of significant peptides containing enriched cationic and aromatic amino acids warrants further investigation. The major question surrounding these is whether or not this arises from antibodies directed towards a particular antigen(s), or if it is a nonspecific consequence of being infected. This could be shown with a simple experiment which removes disease-specific antibodies by treating infected sera with pathogen-coated magnetic beads. If charged amino acid containing peptides drop in intensity after purification, then this is a pathogen specific effect. Otherwise, there is likely some other mechanism causing these enrichments. This would have to be repeated with several samples and several diseases, as the effects observed here are somewhat weak, requiring several samples. It is also interesting to note that this effect was observed in the Dog dataset, which consisted of cancers and normal samples **(Figure 5.12)**, only here it was anionic residues rather than cationic ones which were enriched.

Very little modeling has been done on peptide microarrays, and these results indicate that even the simplest variables have explanatory power. If immunosignatures are ever to be understood on a system-wide mechanistic level, quantitative statistical models are the way forward. The next step is explaining within-array binding in terms of motifs or subsequences rather than single amino acids, then extending this to include multiple samples from different diseases to explain between-array and between-class variance. The models tested here are a first step, but they are very basic and deserve more attention to be brought into practical use.

Infections - HT22 - Downs

seq	n_enriched	n_library	p_binom	frac	p_adjusted
A	10891	129092	1.11E-16	0.084366	1.00E-15
G	32179	415597	1.11E-16	0.077428	1.20E-15
K	12278	100615	1.11E-16	0.12203	1.50E-15
P	12867	146545	1.11E-16	0.087802	2.00E-15
S	17414	210675	1.11E-16	0.082658	3.00E-15
R	14293	130838	1.11E-16	0.109242	6.01E-15
V	8987	125975	0.660673	0.07134	1
Q	4270	60357	0.797064	0.070746	1
N	6067	93250	1	0.065062	1
E	6682	137037	1	0.048761	1
D	6018	114500	1	0.052559	1
F	3407	70925	1	0.048037	1
H	6839	129910	1	0.052644	1
L	6090	102543	1	0.05939	1
W	2932	113671	1	0.025794	1
Y	4664	94086	1	0.049572	1

Infections - HT22 - Ups

seq	n_enriched	n_library	p_binom	frac	p_adjusted
F	7394	108827	1.11E-16	0.067943	1.50E-15
K	10595	128429	1.11E-16	0.082497	2.00E-15
L	7979	130167	1.11E-16	0.061298	3.00E-15
S	18996	292848	1.11E-16	0.064866	6.01E-15
Y	7475	135275	3.55E-12	0.055258	3.84E-11
A	8630	162883	0.000279	0.052983	0.002512
G	27936	542251	0.064068	0.051519	0.495077
Q	4141	80003	0.198274	0.051761	1
R	8117	158223	0.354876	0.051301	1
P	8351	162821	0.360757	0.051289	1
N	6003	123086	0.999898	0.048771	1
V	6962	145097	1	0.047982	1
E	2587	147001	1	0.017599	1
D	3367	134464	1	0.02504	1
H	8439	183676	1	0.045945	1
W	7131	184997	1	0.038547	1

Dogs - CIM10Kv3 - Downs

seq	n_enriched	n_library	p_binom	frac	p_adjusted
D	156	2624	0.000442	0.059451	0.031836
E	152	2653	0.00218	0.057294	0.078424
P	134	2358	0.004859	0.056828	0.116534
Q	136	2617	0.055	0.051968	0.989373
T	118	2293	0.082609	0.051461	1
G	258	5315	0.135252	0.048542	1
S	256	5337	0.183831	0.047967	1
A	110	2302	0.2843	0.047785	1
N	113	2381	0.307425	0.047459	1
K	109	2323	0.353454	0.046922	1
C	134	2941	0.477209	0.045563	1
R	93	2123	0.625899	0.043806	1
H	104	2521	0.838992	0.041253	1
V	89	2179	0.841942	0.040844	1
L	92	2275	0.869467	0.04044	1
M	93	2301	0.872158	0.040417	1
I	78	2001	0.914229	0.038981	1
W	66	2010	0.997334	0.032836	1
F	71	2252	0.999508	0.031528	1
Y	146	4237	0.999853	0.034458	1

Dogs - CIM10Kv3 - Ups

seq	n_enriched	n_library	p_binom	frac	p_adjusted
L	17	5610	0.004177	0.00303	0.30054
V	15	5503	0.016069	0.002726	0.578122
R	14	5756	0.04428	0.002432	1
C	13	6989	0.229265	0.00186	1
T	9	5237	0.327947	0.001719	1
P	9	5288	0.338655	0.001702	1
A	9	5398	0.361958	0.001667	1
F	10	6112	0.386642	0.001636	1
I	9	5556	0.395772	0.00162	1
K	9	5658	0.417707	0.001591	1
H	9	5958	0.48199	0.001511	1
G	19	12294	0.500212	0.001545	1
E	7	4809	0.503677	0.001456	1
S	18	12357	0.603837	0.001457	1
N	6	4897	0.66891	0.001225	1
Q	6	5024	0.694535	0.001194	1
M	6	5155	0.7196	0.001164	1
W	6	6178	0.867522	0.000971	1
Y	14	12093	0.877836	0.001158	1
D	4	4937	0.898153	0.00081	1

Dengue - CIM10Kv2 - Downs

seq	n_enriched	n_library	p_binom	frac	p_adjusted
K	924	4053	1.11E-16	0.227979	7.99E-15
H	789	4066	1.01E-08	0.194048	3.63E-07
D	735	3761	1.13E-08	0.195427	2.72E-07
Q	681	3748	0.000201	0.181697	0.00362
S	1384	8031	0.000775	0.172332	0.011147
M	669	3781	0.002313	0.176937	0.027744
P	530	2956	0.002454	0.179296	0.025223
G	1327	7780	0.003365	0.170566	0.030264
E	663	3874	0.024548	0.171141	0.19626
N	594	3544	0.088116	0.167607	0.634035
A	596	3617	0.174553	0.164777	1
C	687	4311	0.455556	0.15936	1
R	482	3107	0.691926	0.155134	1
T	556	3617	0.778577	0.153719	1
V	494	3673	0.999942	0.134495	1
L	449	3597	1	0.124826	1
W	403	3379	1	0.119266	1
F	392	3661	1	0.107075	1
I	391	3692	1	0.105905	1
Y	734	6612	1	0.11101	1

Dengue - CIM10Kv2 - Ups

seq	n_enriched	n_library	p_binom	frac	p_adjusted
P	1019	5119	3.97E-09	0.199062	2.85E-07
Y	2006	10694	8.58E-09	0.187582	3.09E-07
R	1053	5333	9.84E-09	0.19745	2.36E-07
W	1031	5292	1.40E-07	0.194822	2.52E-06
H	967	4988	7.21E-07	0.193865	1.04E-05
G	1735	10305	0.274023	0.168365	1
K	756	4506	0.381503	0.167776	1
C	941	5628	0.411142	0.1672	1
M	681	4123	0.550842	0.165171	1
F	844	5175	0.700919	0.163092	1
S	1655	10152	0.785457	0.163022	1
T	692	4399	0.926707	0.157308	1
A	706	4579	0.978532	0.154182	1
L	720	4785	0.996874	0.15047	1
Q	653	4383	0.997911	0.148985	1
N	624	4200	0.997947	0.148571	1
D	619	4261	0.999722	0.145271	1
I	605	4350	0.999997	0.13908	1
V	606	4443	1	0.136394	1
E	487	4045	1	0.120396	1

Table 5.4: Enrichment tables for three immunosignature datasets on diverse platforms: These tables list enrichment scores for each amino acid. The column labels are as follows: *seq* – amino acid, *n_enriched* – amino acid counts in significant peptides, *n_library* – amino acid counts in the entire array, *p_binom* – binomial approximation of the p-value for enrichment, *frac* – fraction of the total counts in the significant peptides, *p_adjusted* – FDR corrected p-value. Down refers to those peptides lower in intensity on average in disease as compared to normals, and up refers to those that are higher in intensity in disease relative to normals. Charged amino acids were enriched in down peptides in all three datasets, while aromatics predominated in the up peptides.

CHAPTER 6

PROTOMAPPER: SOFTWARE FOR RAPID DISCOVERY OF MOTIFS

Preface

This chapter contains a number of terms that may be unfamiliar to the uninitiated, but nonetheless precise. It is important to understand that these terms are commonly used and not limited to one domain. They are about as domain specific as “Microsoft Word”, “Western Blot”, or “ELISA”. As previous chapters have assumed the reader is familiar with these terms, this chapter assumes the reader is familiar with the following terms.

Regular Expressions

Regular Expressions are a concise language used to represent patterns in strings. For example, the regular expression DOGS|CATS would match the following two strings:

1. DOGS
2. CATS

However, if we expand this regular expression as follows, we can make it more flexible. Consider the expression .*(DOGS|CATS).*. The dot (.) means match any character in the alphabet, and the star (*) means that the preceding character can be repeated any number of times. So this expression would match an infinite number of strings containing either the words DOGS or CATS

1. I like DOGS
2. I like CATS
3. I like DOGS but not CATS

…

Regular expressions are flexible and can be combined in complex ways to match many different types of strings. The basic concepts are well stated by Wikipedia: en.wikipedia.org/wiki/Regular_expression.

In general the "building blocks" of regular expressions consist of the following:

Boolean "or"

grey|Grey|gray matches either the words grey, Grey, or gray

Grouping

Parentheses can be used to nest, or group regular expressions. For example (AB)* would match

AB
ABAB
ABABAB
…

Quantification

Quantifiers act upon the preceding character or group. Here are some examples:

? – match zero or one of the preceding element

* - match zero or more of the preceding element

+ - match 1 or more of the preceding element

There are other quantifiers not covered here.

Finite Regular Expression

A finite regular expression is any regular expression that matches a finite number of strings. Many regular expressions match an infinite number of strings (for example any expression containing a *). This chapter concerns itself with a method for rapidly matching finite regular expressions.

Grep

Grep is mentioned several times in this chapter. This is a well known program used to match regular expressions on Unix based systems. It is so commonly used that there are grep functions in many major programming languages such as R and Python. This is a highly optimized O(n) algorithm and software for matching regular expressions. Again, Wikipedia contains an excellent introduction to this program: http://en.wikipedia.org/wiki/Grep.

Big O Notation

Big O notation is used to state the time complexity of an algorithm relative to its input. The time it takes an O(n) algorithm to complete will increase approximately linearly as the size of its input increases. Similarly, an $O(n^2)$ algorithm time will increase as a function of the square of its input. Again, Wikipedia contains an excellent introduction http://en.wikipedia.org/wiki/Big_O_notation.

N-Gram

An N-gram is simply a sequence of characters of length N. For example, a trigram is any string of length 3. The language of all trigrams can be described with the following regular expression: (…). This chapter uses the word trigram often to describe an indexing strategy. This means that trigrams are used to rapidly retrieve documents containing that trigram.

Naïve

Naïve, as used in this chapter, means "the most obvious method available". Here it is used to refer to those methods that do not use extra information in the form of an index to match (finite) regular expressions.

The terms defined above are important for understanding what this chapter is about. This chapter is entirely concerned with a strategy for rapid matching of regular expressions, which is useful for, but not limited to, certain biological applications. The primary result is that this strategy is much faster than grep or any other naive method for certain subclasses of regular expressions.

Abstract

Biologists have been using regular expressions to represent and search for motifs for many years. Databases such as ProSite gather patterns representing common biological motifs, but tools for searching these patterns against the known sequence information are naïve methods. Though efficient and O(n), this is not good enough for the current size of the database if one is to run a web server at reasonable cost, a problem that will only continue to worsen. In this manuscript we test a trigram indexing strategy based on Apache Lucene that can resolve low complexity finite regular expressions at speeds that are logs faster than even the most optimized naïve methods. Benchmarking against grep (a standard and highly optimized O(n) method for regular expression matching) was conducted, and patterns requiring fewer than 6025 index lookups resulted in search times much faster than grep. With further optimization, this method would be good replacement for naïve systems such as ScanProSite which are showing their inefficiency under the growing size of the publically available sequence information. Source code is available under the BSD License at https://github.com/joshuaar/Protomapper-Search.

Introduction

Short, linear patterns in peptide sequences are highly common in nature, including linear antibody epitopes (Buus et al., 2012a), conserved signal peptides (Chaddock et al., 1995), and peptide-protein binding motifs (Bähler & Rhoads, 2002). Locating specific patterns in a database is useful in proteomic studies, immunological epitope identification, and sequence based protein retrieval. BLAST, a commonly used technique for finding similar sequences, provides limited capability for regular expression pattern searches, but is not ideal for patterns with variable lengths or with position specific

substitution allowances. The Protein Information Resource (a website for searching biological databases) has recently retired its pattern matching service, and ScanProsite (Gattiker, Gasteiger, & Bairoch, 2002), while useful and flexible, uses naïve methodology and was developed over a decade ago when the amount of publically available sequence information was much lower. Today UniProtKB contains over 300GB of sequences, which are not easily searchable using standard regular expression search utilities such as grep or previously developed bioinformatics tools. More recently, a search utility for rapid, exact peptide matches has been developed using similar technology (C. Chen et al., 2013), but fast regular expression based queries remain elusive. This paper describes, tests and provides a reference implementation for an indexing strategy and method of fast regular expression matching.

While regular expression matching fundamentally has linear time complexity (Sipser, 1996), in practice, performant search over large databases is challenging. None of the major SQL databases provide strong support for regular expression searching over long strings of text (such as peptide sequences), and popular "out of the box" solutions such as Solr and Apache Lucene are optimized for search over human readable text consisting of short words separated by whitespace. A new, scalable and general approach is needed, and this paper develops one such method that could meet this need for a rapid engine matching a finite subset of the regular expressions. We have made this tool available under the BSD license (https://github.com/joshuaar/Protomapper-Search). This method draws inspiration from Cho et. al. (Cho et al., 2003), Google Code Search (Cox, 2012), and Chen et.al. (Chen et al., 2013), consisting of an expression compiler which produces Lucene queries to a trigram index of sequences.

Cho was the first recorded regular expression indexing engine. It analyzes the expression to find a number of trigram lookups, and then returns a superset of documents for additional matching. Google Code Search employed a similar strategy, but included optimizations for special cases, taking special care to analyze prefixes and suffixes of the expression to construct optimal queries. Chen created a peptide search engine using Lucene, which only matches exact sequences. The method discussed here combines the method from Chen with those from Cho and Google Code Search.

Methods

Protomapper uses two underlying databases. The first is from NCBI's FTP site, and includes proteomes of bacteria and viruses organized by strain. The second is UniprotKB, which contains over 40 million sequences, many redundant. These databases are indexed using Apache Lucene, and fields include the peptide sequence, accession number, protein description, and organism.

The protein sequences are passed through Lucene's NGramTokenFilter. This generates overlapping trigrams of the sequence, allowing the search space to be indexed in terms of trigrams. The structure is that of an inverted index **(Figure 6.2A)** which means that one can look up sequences by trigram at O(1) time complexity. This index structure follows a similar approach to previous methods such as PeptideMatch (3-mers) (Chen et al., 2013), UniPept (6-mers) (Mesuere et al., 2012) and SANS (suffixes) (Koskinen & Holm, 2012), though these methods were not designed for pattern matching.

```
re          -->  choice "^" re | choice
choice      -->  term "|" choice | term
term        -->  factor | factor * term
factor      -->  base | base * lenRange
lenRange    -->  {\d+,\d+}
base        -->  "(" re ")" | "[" range "]" | character
range       -->  compliment * subRanges
compliment  -->  empty | "^"
subRanges   -->  subRange | subRange * subRanges
subRange    -->  character | character "-" character
character   -->  \w
```

Figure 6.1: Context Free Grammar Recognized by the Query Compiler: This is the language recognized by the query compiler. It is a context free grammar and a subset of the regular expressions. Each string in this language is also a regular expression, but not every regular expression is in this language. This grammar is equivalent to the "finite regular expressions". Finite regular expressions refer to those expressions such as (A[XYZ]C) that match a finite number of strings (as opposed to something like (A.*C) which matches an infinite number of strings).

This index structure is conducive to limited regular expression matching, where the query speed is dependant primarily on the complexity of the pattern, not the number of sequences contained in the database. In order to search the index for regular expression matches, we developed a compiler for converting query strings into Lucene MultiPhraseQuery and BooleanQuery objects, which directly read the index resulting in a list of matching sequence identifiers. The grammar recognized by the compiler a finite subset of the regular expressions and is given in **Figure 6.1**. It allows the user to match finite length wildcards (say, for example any character repeated between 0 and 20 times), variable length gaps and character substitutions at high speed. The compiler goes through three steps. The first two steps modify the query string to remove "or" symbols (|) and length ranges, replacing those with a list of query strings. The final step converts each query string into a Lucene Query which searches the index for matching strings,

completing the compilation process. An example of how a query string is converted into a Lucene query is given in **Figure 6.2B**.

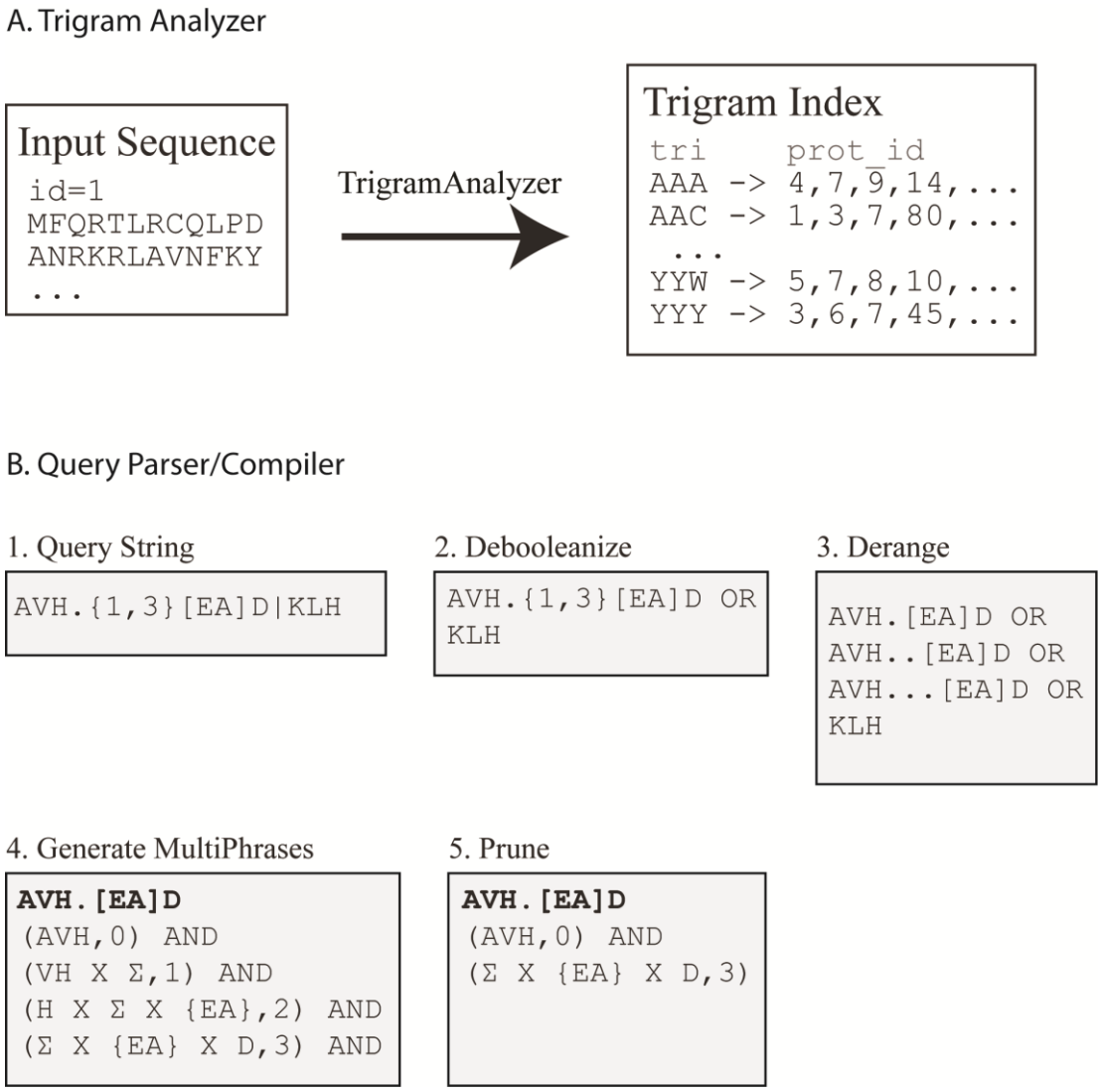


Figure 6.2: Indexing and Query Compilation Procedures: (A): Input sequences are processed and inserted into an inverted trigram index. Each trigram window from the sequence corresponds to a key in the index. Each trigram key maps back to a set of pointers to the sequences containing that trigram. This enables O(1) lookups of any given trigram, and facilitates the construction of more complex queries. (B) Complex finite regular expressions must be transformed into a series of index queries. This requires a multistep procedure where boolean and ranged regular expressions are converted into simpler but equivalent versions. These simplified regular expressions are converted into MultiPhraseQueries which consist entirely of trigram lookups. These queries are then pruned so as to use as few index lookups as possible.

The compiler and web interface are written in Scala, and all communication between the client and the database is done through a simple RESTful API. Details about architecture and API syntax are given in **Figure 6.3**. The index and querying system is built using Apache Lucene 4.0, and the API and interface are served through the Play framework. These two functions are decoupled such that the search methods could be embedded into existing databases and APIs in a flexible way.

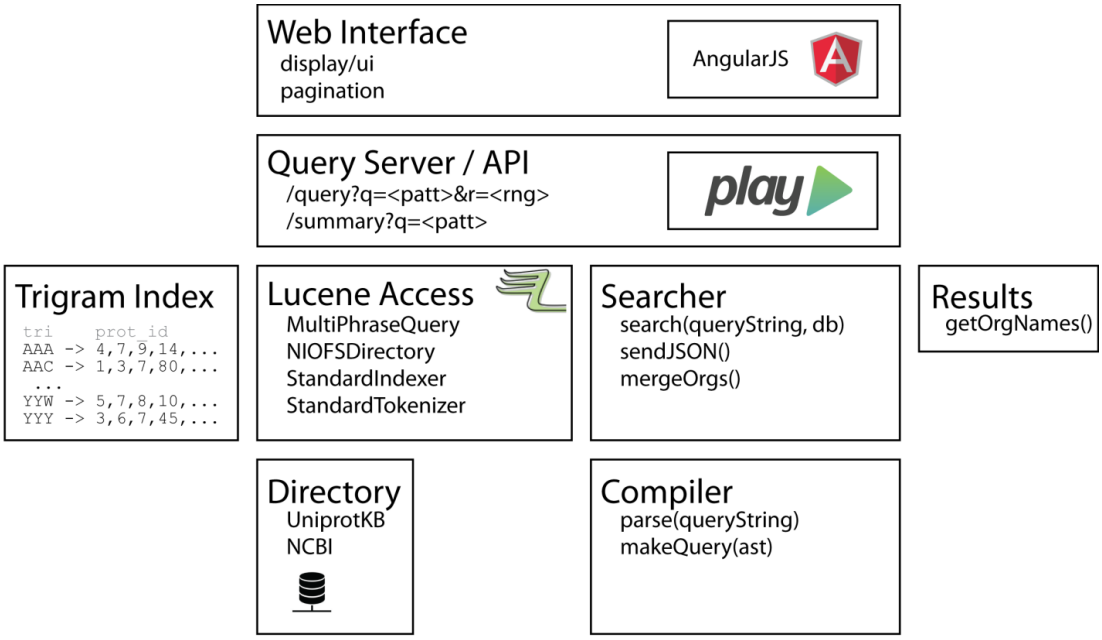


Figure 6.3: Protomapper Architecture: This block diagram shows the various modules within Protomapper and how they interact. The Web interface communicates with the Play API. The API communicates with the search module and also the underlying Lucene Access module in order to compile and process queries. The searcher embeds a compiler to deal with turning user strings into compiled queries. The Lucene access module handles all reading and writing to the Lucene index.

Results

Protomapper is up to three orders of magnitude faster than naïve methods and existing pattern searchers. The catch is that query duration in Protomapper increases exponentially with more complex queries which is a major limitation. Complexity refers in this case not to time complexity, but the number of times the system must access the index for trigram lookups. A complexity score of 1000, for example, corresponds to 1000 Lucene trigram lookups. In Lucene parlance this involves querying 1000 PhraseQuery objects against the index. We compared Protomapper query durations to grep (a naïve method) for varying database sizes and query complexities. For both search methods, there is a clear log-log linear relationship between query complexity and duration. For low complexity queries, Protomapper is several logs faster than grep, but at higher complexities the situation is reversed. The inflection point remained relatively constant for each database size tested, so it would be relatively straightforward to design a query optimizer around the complexity score. In all cases, the inflection point was at around log(complexity)=3.78, which corresponds to 6025 trigram lookups. Protomapper is an effective strategy for queries with fewer lookups than this. A table of example queries and their corresponding complexities is given in **Table 6.1**.

Complexity	Log-Complexity	Expression
3	0.477121255	AVHAD
4	0.602059991	SNKQRLP
5	0.698970004	KQRLSGGGGGGG
9	0.954242509	DFKHKWLAAAAARRRRLLLLPPPP
17	1.230448921	[IL][RQ][TC][MR][MK]
24	1.380211242	AVHAD..DDR
31	1.491361694	{10,10}A{10,10}R{10,10}L{10,10}M{10,10}O{10,10}P
36	1.556302501	YK[DE][SG]TLI[IML]QL[LF][RHC]D [LF]T[LS]W[TANS][SAD]
64	1.806179974	A.D.R.LL.P
66	1.819543936	DDR.{2,23}.HRT
89	1.949390007	G[FYIL][DE][LIVMT][DE][LIVMF] PSYG[LIVMA][VAGC]TPRGGL[LIVMAGN]
152	2.181843588	{2,2}(G[FYIL][DE][LIVMT][DE][LIVMF]PSYG[LIVMA] [VAGC]TPRGGL[LIVMAGN])
197	2.294466226	[EQ][LNYH].[ATV][FY][LDAM][T]W[PG]N
253	2.403120521	[IL].[RQ].[TC].[MR].[MK]
421	2.624282096	[IL].[RQ].[TC].[MR].[MK][IL].[RQ].[TC].[MR].[MK]
527	2.721810615	C.PC..CCP..C[PEG]
2131	3.32858345	[LIVM].[SADN]..C.R[LIVM]....[GSC]H[STA]
2773	3.44294987	C...C..[LMF]...[DEN][LI].....C
3655	3.562887381	[FILV]Q...[RK]G...[RK]..[FILVWY]
4684	3.670616886	C...C..[LMF]...[DEN][LI].....CC...C..[LMF]...[DEN][LI].....C
5589	3.74733411	[LIVMTR].[LIVMT][LIVMF].[GATMC][ST][NS] ....[LIVM]D..[AS][LIFAV].{1,3}R
16444	4.216007468	[TG][STV].....[LIVMF]..R...[DEQNH]..S....[IFY] .....[LIVMF]...[LIVMF].....I.....[LIVMFA]..[LIVMF]
30586	4.485522684	{2,2}([TG][STV].....[LIVMF]..R...[DEQNH]..S ....[IFY].....[LIVMF]...[LIVMF].....I.....[LIVMFA].. [LIVMF])

Table 6.1: Various Queries and their Complexity Scores: These queries are ordered to increasing complexity. Complexity refers to the number of trigram lookups required to resolve the pattern within the index. A log complexity score between 3 and 3.5 takes approximately equal time whether the index or naïve methods are used. Log complexities above 3.5 are slower using the index, and scores below 3 are faster using the index. **Figure 6.4** summarizes the benchmarking tests giving rise to these cutoffs. The complexity score would be the basis of a query optimization system, whereby different methods could be selected based on analysis of the query.

Since the index is constructed on trimers, queries that contain many gaps run slower since many possible trimers could occur at the gapped position. Also, variable length queries using the x{a,b} syntax are compiled to BooleanQuery objects, which also increase the number of index lookups required. An analysis of various queries and their speed is given in **Figure 6.4**.

The user interface allows several options. Searches can be directed towards the entire database, or a specific subset of organisms. Results are returned as a paginated table, and the most prevalent genera are given in a pie chart. Results can be downloaded as a fasta file for storage and later retrieval. We also provide a simple API that returns JSON formatted results, and source code which can be incorporated into existing projects or used as a command line application.

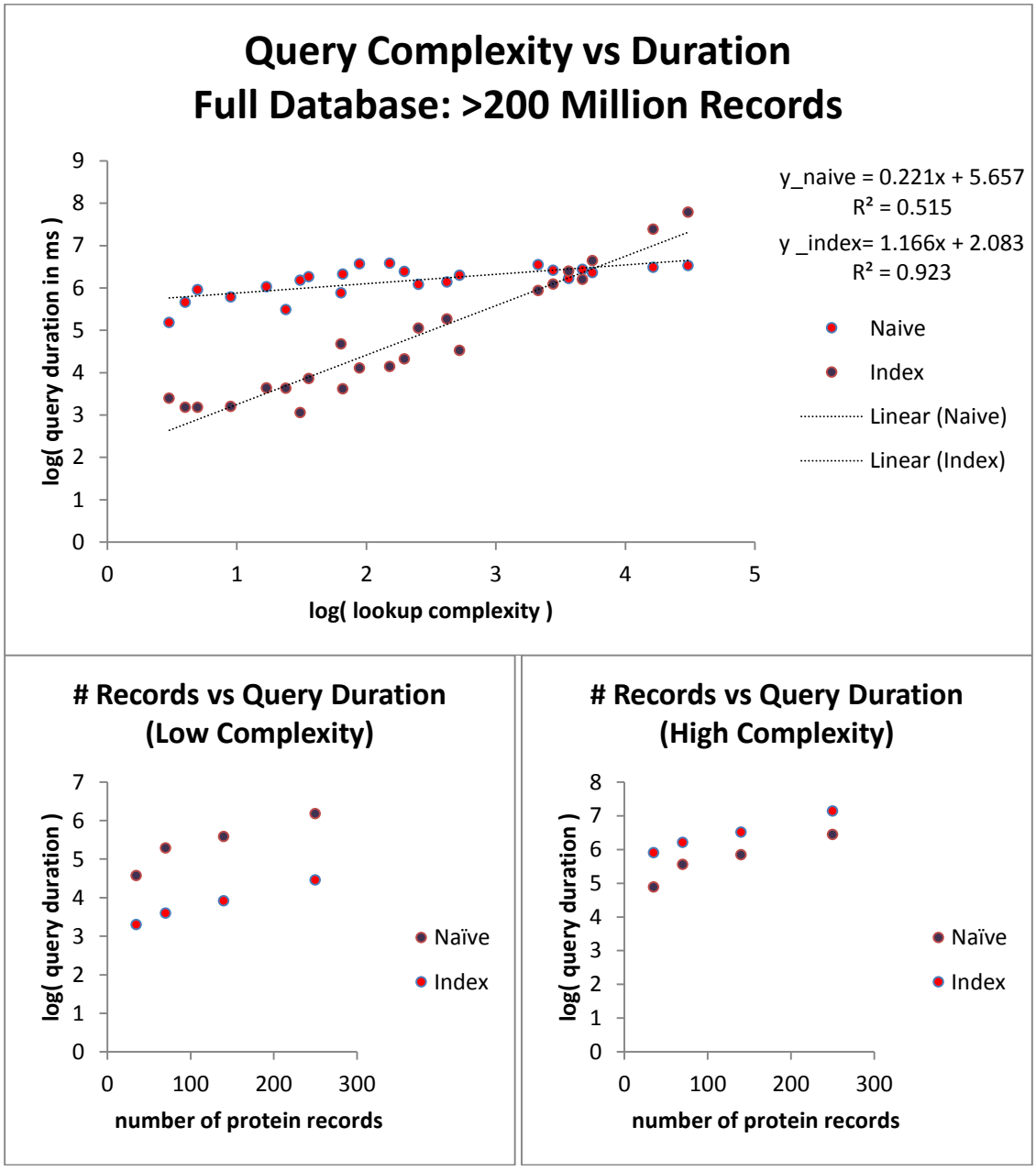


Figure 6.4: Protomapper vs. Grep (naïve method): Benchmark tests comparing the Protomapper index versus grep. Grep is a mature and optimized linear method for matching regular expressions, and is a standard utility in most linux distributions. For low complexity queries (less than 6025 index lookups) Protomapper is logs faster than grep. However, speed decreases exponentially with increasing index lookups, causing very poor performance for high complexity queries. In such cases, a query optimizer should search using naïve methods. Both grep and protomapper search more slowly in large databases, affected more severely.

Discussion

An indexing strategy was designed in order to rapidly search peptide motifs across a large database of protein data. This employed a trigram index and the Lucene search library along with a pattern compiler. Motifs are encoded as finite regular expressions and compiled into queries which can be directly executed on Lucene's engine. The system was benchmarked against the naïve pattern matcher grep, which is the standard and most optimized method for these types of searches. Finally a web and command line interface named Protomapper was constructed to allow easy access to the system. Protomapper is fast for low complexity patterns, but speed decreases exponentially with increased pattern complexity.

These results (**Table 6.1**) highlight an important issue surrounding these types of indexing strategies. Only a few strategies exist in the literature (Cho & Rajagopalan, 2002; Cox, 2012), and they all run into the same fundamental flaw: regular expressions are arbitrarily complex, so there is no indexing strategy that is guaranteed to be faster than naïve methods (which are, admittedly, highly mature and optimized). The best that can be achieved is a system optimized for some portion of the regular expressions. In this method, un-nested, finite regular expressions were chosen as the language **(Figure 6.1** for precise generative definition**)**. This guarantees at least that patterns can be matched solely through interrogation of the index, and makes the problem easier to analyze. The second restriction is that patterns fall below a certain complexity threshold, guaranteeing the impossibility of querying intractable patterns.

In practical terms, this means that this system would be useful for looking up any peptide containing less than 6025 non-overlapping trigrams. Any peptide less than 18,075

amino acids in length would result in efficient lookup. Of course, this system does not just match exact sequences, but patterns as defined by the language in **Figure 6.1**. This creates a much more complex situation as to which patterns would be useful under this system. **Table 6.1** contains a number of patterns color coded by whether lookup time would be faster than naïve methods using this system.

There are several optimizations that could be employed in order to accommodate high complexity patterns. Currently, the compiler resolves the expression entirely through index lookups. That is, after the index is interrogated, the results returned exactly match the regular expression and there is no need for further processing. While this works well when the number of index lookups is low (<6025), it quickly becomes intractable for queries that require more lookups. However, for most patterns, a reasonable superset of results could be returned using far fewer lookups. A strategy might be to truncate the pattern such that fewer than 6025 lookups are employed. Given the nature of finite regular expressions, this method would return a superset that contains a small fraction of the original database, but does not exactly match the pattern. Then, on this restricted subset the naïve method (such as grep) could be employed resulting in an overall reduction in query duration as compared to the full naïve method.

The other indexing strategies mentioned (J. Cho & Rajagopalan, 2002; Cox, 2012) are very similar to this one. The distinguishing factor of Protomapper is that it is built on top of Lucene and its patterns can be resolved using only index lookups. The other difference is that Protomapper restricts itself to only a portion of the regular expressions which simplifies the problem without sacrificing too much expressive power for its intended application of motif searching. With further optimization this would be an

excellent candidate to replace tools such as ScanProSite. The source code is available for download and modification under the BSD licence.

https://github.com/joshuaar/Protomapper-Search

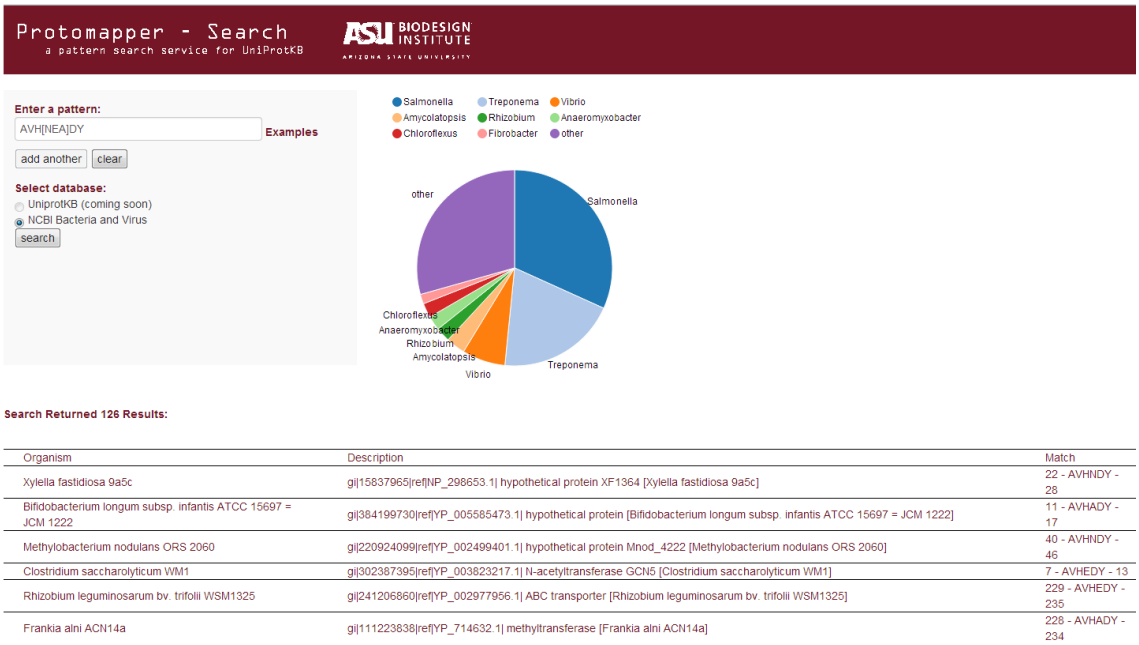


Figure 6.5: User Interface Screenshot: Users select a database, add multiple patterns and search. The distribution of organisms is shown in a pie graph, and results are displayed in a table below. These can be downloaded in fasta format. The program is also available as a command line application for use in server-side procedures.

CHAPTER 7

CONCLUDING REMARKS

Immunosignatures are incredibly complex assays that still are barely understood. This issue extends to array-based and high-throughput assays in general where scientists struggle to apply their expertise in single experiments and multiply these thousands of times. It is clear that this transition is not straightforward and gives rise to unexpected statistical, computational and physical concerns. This thesis reviewed and pointed out some of these difficulties, as well as providing some solutions to some of them using the immense computational power at the disposal of the modern scientist. In the future, these types of approaches will become even more important as yet more of the scientific process is automated with machines. We are in a transition period, and this is glaringly apparent in medical and biological research where sequencing, IT infrastructure, and high throughput assays are changing the way discoveries are made, published, and validated. It is important to embrace this model, where problem finding is just as important as problem solving and requires the same level of investment. Hypothesis driven science still rules, but the science of finding hypotheses and testing them on the existing data is still in its nascent phase.

REFERENCES

Anders, R. (1986). Multiple cross-reactivities amongst antigens of Plasmodium falciparum impair the development of protective immunity against malaria. *Parasite immunology, 8*(6), 529-539.

Arnon, R., Tarrab-Hazdai, R., & Steward, M. (2000). A mimotope peptide-based vaccine against Schistosoma mansoni: synthesis and characterization. *Immunology, 101*(4), 555-562. doi: 10.1046/j.1365-2567.2000.00139.x

Bähler, M., & Rhoads, A. (2002). Calmodulin signaling via the IQ motif. *FEBS Letters, 513*(1), 107-113. doi: http://dx.doi.org/10.1016/S0014-5793(01)03239-2

Ballew, J. T., Murray, J. A., Collin, P., Mäki, M., Kagnoff, M. F., Kaukinen, K., & Daugherty, P. S. (2013). Antibody biomarker discovery through in vitro directed evolution of consensus recognition epitopes. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1314792110

Banoo, S., Bell, D., Bossuyt, P., Herring, A., Mabey, D., Poole, F., . . . Linke, R. (2008). Evaluation of diagnostic tests for infectious diseases: general principles. *Nature Reviews Microbiology, 8*, S16-S28.

Brown, G., Culvenor, J., Crewther, P., Bianco, A., Coppel, R., Saint, R., . . . Anders, R. (1985). Localization of the ring-infected erythrocyte surface antigen (RESA) of Plasmodium falciparum in merozoites and ring-infected erythrocytes. *The Journal of experimental medicine, 162*(2), 774-779.

Brown, J. R., Stafford, P., Johnston, S. A., & Dinu, V. (2011). Statistical methods for analyzing immunosignatures. *BMC bioinformatics, 12*(1), 349.

Buus, S., Rockberg, J., Forsström, B., Nilsson, P., Uhlen, M., & Schafer-Nielsen, C. (2012a). High-resolution Mapping of Linear Antibody Epitopes Using Ultra high-density Peptide Microarrays. *Molecular & Cellular Proteomics, 11*(12), 1790-1800. doi: 10.1074/mcp.M112.020800

Buus, S., Rockberg, J., Forsström, B., Nilsson, P., Uhlen, M., & Schafer-Nielsen, C. (2012b). High-resolution Mapping of Linear Antibody Epitopes Using Ultrahigh-density Peptide Microarrays. *Molecular & Cellular Proteomics, 11*(12), 1790-1800. doi: 10.1074/mcp.M112.020800

Campo, D. S., Dimitrova, Z., Yokosawa, J., Hoang, D., Perez, N. O., Ramachandran, S., & Khudyakov, Y. (2012). Hepatitis C Virus Antigenic Convergence. [10.1038/srep00267]. *Sci. Rep., 2*. doi: http://www.nature.com/srep/2012/120215/srep00267/abs/srep00267.html#supplementary-information

Chaddock, A. M., Mant, A., Karnauchov, I., Brink, S., Herrmann, R. G., Klösgen, R., & Robinson, C. (1995). A new type of signal peptide: central role of a twin-arginine motif in transfer signals for the delta pH-dependent thylakoidal protein translocase. *The EMBO journal, 14*(12), 2715.

Chen, C., Li, Z., Huang, H., Suzek, B. E., Wu, C. H., & Consortium, U. (2013). A fast Peptide Match service for UniProt Knowledgebase. *Bioinformatics, 29*(21), 2808-2809. doi: 10.1093/bioinformatics/btt484

Chen, Y., Pan, Y., Guo, Y., Qiu, L., Ding, X., & Che, X. (2010). Comprehensive mapping of immunodominant and conserved serotype- and group-specific B-cell epitopes of nonstructural protein 1 from dengue virus type 1. *Virology, 398*(2), 290-298. doi: http://dx.doi.org/10.1016/j.virol.2009.12.010

Chen, Z., & Gu, J. (2007). Immunoglobulin G expression in carcinomas and cancer cell lines. *The FASEB Journal, 21*(11), 2931-2938. doi: 10.1096/fj.07-8073com

Cho, H.-S., Mason, K., Ramyar, K. X., Stanley, A. M., Gabelli, S. B., Denney, D. W., & Leahy, D. J. (2003). Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. [10.1038/nature01392]. *Nature, 421*(6924), 756-760. doi: http://www.nature.com/nature/journal/v421/n6924/suppinfo/nature01392_S1.html

Cho, J., & Rajagopalan, S. (2002). *A fast regular expression indexing engine.* Paper presented at the 2013 IEEE 29th International Conference on Data Engineering (ICDE).

Cox, R. (2012). Regular Expression Matching with a Trigram Index or How Google Code Search Worked Retrieved 1/1/2014, from http://swtch.com/~rsc/regexp/regexp4.html

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research, 14*(6), 1188-1190.

. DENV DetectTM IgM CAPTURE ELISA. (2012), 2014, from http://www.inbios.com/elisas/denv-detect-igm-ELISA

Diehnelt, C. W., Shah, M., Gupta, N., Belcher, P. E., Greving, M. P., Stafford, P., & Johnston, S. A. (2010). Discovery of High-Affinity Protein Binding Ligands - Backwards. *PLoS One, 5*(5), e10728. doi: 10.1371/journal.pone.0010728

Dunphy, E. J., & McNeel, D. G. (2005). Antigen-specific IgG elicited in subjects with prostate cancer treated with flt3 ligand. *Journal of immunotherapy, 28*(3), 268-275.

Edfors, F., Boström, T., Forsström, B., Zeiler, M., Johansson, H., Lundberg, E., . . . Uhlen, M. (2014). Immunoproteomics Using Polyclonal Antibodies and Stable Isotope-labeled Affinity-purified Recombinant Proteins. *Molecular & Cellular Proteomics, 13*(6), 1611-1624. doi: 10.1074/mcp.M113.034140

Fack, F., Hügle-Dörr, B., Song, D., Queitsch, I., Petersen, G., & Bautz, E. K. (1997). Epitope mapping by phage display: random versus gene-fragment libraries. *Journal of Immunological Methods, 206*(1), 43-52.

Foley, M., Tilley, L., Sawyer, W. H., & Anders, R. F. (1991). The ring-infected erythrocyte surface antigen of< i> Plasmodium falciparum</i> associates with spectrin in the erythrocyte membrane. *Molecular and biochemical parasitology, 46*(1), 137-147.

Forsström, B., Axnäs, B. B., Stengele, K.-P., Bühler, J., Albert, T. J., Richmond, T. A., . . . Uhlen, M. (2014). Proteome-wide Epitope Mapping of Antibodies Using Ultra-dense Peptide Arrays. *Molecular & Cellular Proteomics, 13*(6), 1585-1597. doi: 10.1074/mcp.M113.033308

Frank, S. (2002). *Immunology and Evolution of Infectious Disease*: Princeton University Press.

Garcia G, V. D., Del Angel RM. (1997). Recognition of synthetic oligopeptides from nonstructural proteins NS1 and NS3 of dengue-4 virus by sera from dengue virus-infected children. *Am J Trop Med Hyg, 56*(4), 466-470.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., . . . Barrell, B. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. [10.1038/nature01097]. *Nature, 419*(6906), 498-511. doi: http://www.nature.com/nature/journal/v419/n6906/suppinfo/nature01097_S1.html

Gattiker, A., Gasteiger, E., & Bairoch, A. (2002). ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied bioinformatics, 1*(2), 107-108.

Greiff, V., Redestig, H., Luck, J., Bruni, N., Valai, A., Hartmann, S., . . . Or-Guil, M. (2012). A minimal model of peptide binding predicts ensemble properties of serum antibodies. *BMC Genomics, 13*(1), 79.

Greving, M. P., Belcher, P. E., Cox, C. D., Daniel, D., Diehnelt, C. W., & Woodbury, N. W. (2010). High-throughput screening in two dimensions: Binding intensity and off-rate on a peptide microarray. *Analytical biochemistry, 402*(1), 93-95.

Greving, M. P., Belcher, P. E., Diehnelt, C. W., Gonzalez-Moa, M. J., Emery, J., Fu, J., . . . Woodbury, N. W. (2010). Thermodynamic Additivity of Sequence Variations: An Algorithm for Creating High Affinity Peptides Without Large Libraries or Structural Information. *PLoS One, 5*(11), e15432. doi: 10.1371/journal.pone.0015432

Gross, C. P., Anderson, G. F., & Powe, N. R. (1999). The relation between funding by the National Institutes of Health and the burden of disease. *New England Journal of Medicine, 340*(24), 1881-1887.

Guzmán, M. a. G., & Kourí, G. (2004). Dengue diagnosis, advances and challenges. *International journal of infectious diseases, 8*(2), 69-80.

Halperin, R. F. (2011). *Characterization and Analysis of a Novel Platform for Profiling the Antibody Response.* PhD, Arizona State University, ProQuest. Retrieved from http://hdl.handle.net/2286/3v54p6gchnr

Halperin, R. F., Stafford, P., & Johnston, S. A. (2011). Exploring Antibody Recognition of Sequence Space through Random-Sequence Peptide Microarrays. *Molecular & Cellular Proteomics, 10*(3). doi: 10.1074/mcp.M110.000786

Hansen, L. B., Buus, S., & Schafer-Nielsen, C. (2013). Identification and Mapping of Linear Antibody Epitopes in Human Serum Albumin Using High-Density Peptide Arrays. *PLoS One, 8*(7), e68902. doi: 10.1371/journal.pone.0068902

Hansen, M. H., Ostenstad, B., & Sioud, M. (2001). Antigen-specific IgG antibodies in stage IV long-time survival breast cancer patients. *Molecular medicine (Cambridge, Mass.), 7*(4), 230-239.

Hori, S. S., & Gambhir, S. S. (2011). Mathematical Model Identifies Blood Biomarker-Based Early Cancer Detection Strategies and Limitations. *Science Translational Medicine, 3*(109), 109ra116. doi: 10.1126/scitranslmed.3003110

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research, 37*(1), 1-13.

Huang, J., Ofek, G., Laub, L., Louder, M. K., Doria-Rose, N. A., Longo, N. S., . . . Connors, M. (2012). Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. [10.1038/nature11544]. *Nature, advance online publication*. doi: http://www.nature.com/nature/journal/vnfv/ncurrent/abs/nature11544.html#supplementary-information

Hughes, A., Cichacz, Z., Scheck, A. C., Coons, S. W., Johnston, S. A., & Stafford, P. (2012). Immunosignaturing can detect products from molecular markers in brain cancer. *PLoS One, 7*(7), e40201. doi: 10.1371/journal.pone.004020

Hughes, A. K., Cichacz, Z., Scheck, A., Coons, S. W., Johnston, S. A., & Stafford, P. (2012). Immunosignaturing can detect products from molecular markers in brain cancer. *PloS one, 7*(7), e40201.

Hunsperger, E. A., Yoksan, S., Buchy, P., Nguyen, V. C., Sekaran, S. D., Enria, D. A., . . . Drebot, M. (2009). Evaluation of commercially available anti-dengue virus immunoglobulin M tests. *Emerging infectious diseases, 15*(3).

James, C. R., Quinn, J. E., Mullan, P. B., Johnston, P. G., & Harkin, D. P. (2007). BRCA1, a Potential Predictive Biomarker in the Treatment of Breast Cancer. *The Oncologist, 12*(2), 142-150. doi: 10.1634/theoncologist.12-2-142

Karpusas, M., Lucci, J., Ferrant, J., Benjamin, C., Taylor, F. R., Strauch, K., . . . Hsu, Y.-M. (2001). Structure of CD40 Ligand in Complex with the Fab Fragment of a Neutralizing Humanized Antibody. *Structure, 9*(4), 321-329. doi: http://dx.doi.org/10.1016/S0969-2126(01)00590-1

Kliks, S. C., Nisalak, A., Brandt, W. E., Wahl, L., & Burke, D. S. (1989). Antibody-dependent enhancement of dengue virus growth in human monocytes as a risk factor for dengue hemorrhagic fever: DTIC Document.

Klotz, M., Blaes, F., Funke, D., Kalweit, G., Schimrigk, K., & Huwer, H. (1999). Shift in the IgG subclass distribution in patients with lung cancer. *Lung Cancer, 24*(1), 25-30. doi: http://dx.doi.org/10.1016/S0169-5002(99)00014-8

Koskinen, J. P., & Holm, L. (2012). SANS: high-throughput retrieval of protein sequences allowing 50% mismatches. *Bioinformatics, 28*(18), i438-i443. doi: 10.1093/bioinformatics/bts417

Krumpe, L. R. H., Atkinson, A. J., Smythers, G. W., Kandel, A., Schumacher, K. M., McMahon, J. B., . . . Mori, T. (2006). T7 lytic phage-displayed peptide libraries exhibit less sequence bias than M13 filamentous phage-displayed peptide libraries. *PROTEOMICS, 6*(15), 4210-4222. doi: 10.1002/pmic.200500606

Kukreja, M. (2012). *Analysis of Immunosignaturing Case Studies.* PhD Dissertation, Arizona State University, ASU Electronic Dissertations and Theses Retrieved from http://hdl.handle.net/2286/3v54p6gchnr

Kukreja, M., Johnston, S. A., & Stafford, P. (2012). Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases. *Proteomics and Bioinformatics, S6*. doi: doi: 10.4172/jpb.S6-001

Leavy, O. (2010). Therapeutic antibodies: past, present and future. [10.1038/nri2763]. *Nat Rev Immunol, 10*(5), 297-297.

Legutki, J. B., & Johnston, S. A. (2013). Immunosignatures can predict vaccine efficacy. *Proceedings of the National Academy of Sciences, 110*(46), 18614-18619. doi: 10.1073/pnas.1309390110

Legutki, J. B., Magee, D. M., Stafford, P., & Johnston, S. A. (2010). A general method for characterization of humoral immunity induced by a vaccine or infection. [doi: DOI: 10.1016/j.vaccine.2010.04.061]. *Vaccine, 28*(28), 4529-4537.

Legutki, J. B., Zhao, Z.-G., Greving, M., Woodbury, N., Johnston, S. A., & Stafford, P. (2014). Scalable high-density peptide arrays for comprehensive health monitoring. [Article]. *Nat Commun, 5*. doi: 10.1038/ncomms5785

Loscalzo, J. (2006). The NIH budget and the future of biomedical research. *New England Journal of Medicine, 354*(16), 1665-1667.

Luck, K., & Travé, G. (2011). Phage display can select over-hydrophobic sequences that may impair prediction of natural domain‑peptide interactions. *Bioinformatics, 27*(7), 899-902. doi: 10.1093/bioinformatics/btr060

Martin, A. B., Lassman, D., Washington, B., Catlin, A., & Team, t. N. H. E. A. (2012). Growth In US Health Spending Remained Slow In 2010; Health Share Of Gross Domestic Product Was Unchanged From 2009. *Health Affairs, 31*(1), 208-219. doi: 10.1377/hlthaff.2011.1135

Mestre-Ferrandiz, J., Sussex, J., & Towse, A. (2012). The R&D cost of a new medicine. *London: Office of Health Economics (www. fiercebiotech. com/press-releases/new-ohe-study-pharmaceutical-rd-costs-released*.

Mesuere, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., & Dawyndt, P. (2012). Unipept: Tryptic Peptide-Based Biodiversity Analysis of Metaproteome Samples. *Journal of Proteome Research, 11*(12), 5773-5780. doi: 10.1021/pr300576s

Murray, N. E. A., Quam, M. B., & Wilder-Smith, A. (2013). Epidemiology of dengue: past, present and future prospects. *Clinical epidemiology, 5*, 299.

Nagele, E. P., Han, M., Acharya, N. K., DeMarshall, C., Kosciuk, M. C., & Nagele, R. G. (2013). Natural IgG Autoantibodies Are Abundant and Ubiquitous in Human Sera, and Their Number Is Influenced By Age, Gender, and Disease. *PLoS One, 8*(4), e60726. doi: 10.1371/journal.pone.0060726

Namekar, M., Ellis, E. M., O'Connell, M., Elm, J., Gurary, A., Park, S. Y., . . . Nerurkar, V. R. (2013). Evaluation of a New Commercially Available Immunoglobulin M Capture Enzyme-Linked Immunosorbent Assay for Diagnosis of Dengue Virus Infection. *Journal of clinical microbiology, 51*(9), 3102-3106.

Navalkar, K. (2014). *Antibody Based Strategies For Multiplexed Diagnostics* PhD Dissertation, Arizona State University, ProQuest. (3625036)

Navalkar, K. A., Johnston, S. A., Woodbury, N., Galgiani, J., Magee, D. M., Chicacz, Z., & Stafford, P. (2014). "Application of Immunosignatures to diagnosis of Valley Fever". *Clinical and Vaccine Immunology*. doi: 10.1128/cvi.00228-14

Niederfellner, G., Lammens, A., Mundigl, O., Georges, G. J., Schaefer, W., Schwaiger, M., . . . Slootstra, J. W. (2011). Epitope characterization and crystal structure of GA101 provide insights into the molecular basis for type I/II distinction of CD20 antibodies. *Blood, 118*(2), 358-367.

Nobrega, A., Grandien, A., Haury, M., Hecker, L., Malanchère, E., & Coutinho, A. (1998). Functional diversity and clonal frequencies of reactivity in the available antibody repertoire. *European Journal of Immunology, 28*(4), 1204-1215. doi: 10.1002/(sici)1521-4141(199804)28:04<1204::aid-immu1204>3.0.co;2-g

Pammolli, F., Magazzini, L., & Riccaboni, M. (2011). The productivity crisis in pharmaceutical R&D. [10.1038/nrd3405]. *Nat Rev Drug Discov, 10*(6), 428-438.

Paschke, M. (2006). Phage display systems and their applications. *Applied microbiology and biotechnology, 70*(1), 2-11.

Pasternak, N. D., & Dzikowski, R. (2009). PfEMP1: An antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite Plasmodium falciparum. *The International Journal of Biochemistry & Cell Biology, 41*(7), 1463-1466. doi: http://dx.doi.org/10.1016/j.biocel.2008.12.012

Ramachandran, N., Hainsworth, E., Bhullar, B., Eisenstein, S., Rosen, B., Lau, A. Y., . . . LaBaer, J. (2004). Self-assembling protein microarrays. *Science, 305*(5680), 86-90.

Reichert, J. M., & Valge-Archer, V. E. (2007). Development trends for monoclonal antibody cancer therapeutics. [10.1038/nrd2241]. *Nat Rev Drug Discov, 6*(5), 349-356.

Reineke, U. (2004). Antibody Epitope Mapping Using Arrays of Synthetic Peptides. In B. C. Lo (Ed.), *Antibody Engineering* (Vol. 248, pp. 443-463): Humana Press.

Reineke, U., & Sabat, R. (2009). Antibody Epitope Mapping Using SPOT™ Peptide Arrays. In M. Schutkowski & U. Reineke (Eds.), *Epitope Mapping Protocols* (Vol. 524, pp. 145-167): Humana Press.

Restrepo, L., Stafford, P., & Johnston, S. A. (2012). Feasibility of an early Alzheimer's disease immunosignature diagnostic test. *Journal of Neuroimmunology*. doi: 10.1016/j.jneuroim.2012.09.014

Restrepo, L., Stafford, P., Magee, D. M., & Johnston, S. A. (2011a). Application of immunosignatures to the assessment of Alzheimer's disease. *Annals of neurology, 70*(2), 286-295.

Restrepo, L., Stafford, P., Magee, D. M., & Johnston, S. A. (2011b). Application of immunosignatures to the assessment of Alzheimer's disease. *Annals of Neurology*, 5-18. doi: DOI: 10.1002/ana.22405

Riemer, A. B., Kurz, H., Klinger, M., Scheiner, O., Zielinski, C. C., & Jensen-Jarolim, E. (2005). Vaccination With Cetuximab Mimotopes and Biological Properties of Induced Anti‐Epidermal Growth Factor Receptor Antibodies. *Journal of the National Cancer Institute, 97*(22), 1663-1670.

Rigau-Pérez, J. G., Clark, G. G., Gubler, D. J., Reiter, P., Sanders, E. J., & Vance Vorndam, A. (1998). Dengue and dengue haemorrhagic fever. *The Lancet, 352*(9132), 971-977.

Rodi, D. J., Soares, A. S., & Makowski, L. (2002). Quantitative Assessment of Peptide Sequence Diversity in M13 Combinatorial Peptide Phage Display Libraries. *Journal of Molecular Biology, 322*(5), 1039-1052. doi: http://dx.doi.org/10.1016/S0022-2836(02)00844-6

Russell, L. B. (2009). Preventing Chronic Disease: An Important Investment, But Dont Count On Cost Savings. *Health Affairs, 28*(1), 42-45. doi: 10.1377/hlthaff.28.1.42

San Martín, J. L., Brathwaite, O., Zambrano, B., Solórzano, J. O., Bouckenooghe, A., Dayan, G. H., & Guzmán, M. G. (2010). The epidemiology of dengue in the Americas over the last three decades: a worrisome reality. *The American journal of tropical medicine and hygiene, 82*(1), 128-135.

Saphire, E. O., Montero, M., Menendez, A., van Houten, N. E., Irving, M. B., Pantophlet, R., . . . Wilson, I. A. (2007). Structure of a High-affinity "Mimotope" Peptide Bound to HIV-1-neutralizing Antibody b12 Explains its Inability to Elicit gp120 Cross-reactive Antibodies. *Journal of Molecular Biology, 369*(3), 696-709. doi: http://dx.doi.org/10.1016/j.jmb.2007.01.060

Scherf, A., Lopez-Rubio, J. J., & Riviere, L. (2008). Antigenic Variation in Plasmodium falciparum. *Annual Review of Microbiology, 62*(1), 445-470. doi: doi:10.1146/annurev.micro.61.080706.093134

Schwartz, E., Mileguir, F., Grossman, Z., & Mendelson, E. (2000). Evaluation of ELISA-based sero-diagnosis of dengue fever in travelers. *Journal of Clinical Virology, 19*(3), 169-173.

Schweitzer, B., Meng, L., Mattoon, D., & Rai, A. J. (2010). Immune Response Biomarker Profiling Application on ProtoArray® Protein Microarrays *The Urinary Proteome* (pp. 243-252): Springer.

Sharrett, A. R., Ballantyne, C., Coady, S., Heiss, G., Sorlie, P., Catellier, D., & Patsch, W. (2001). Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein (a), apolipoproteins AI and B, and HDL density subfractions the atherosclerosis risk in communities (ARIC) study. *Circulation, 104*(10), 1108-1113.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology, 26*(10), 1135-1145.

Sipser, M. (1996). *Introduction to the Theory of Computation*: International Thomson Publishing.

Sivalingam, G. N., & Shepherd, A. J. (2012). An analysis of B-cell epitope discontinuity. *Molecular Immunology, 51*(3‑4), 304-309. doi: http://dx.doi.org/10.1016/j.molimm.2012.03.030

Smith, G. P., & Petrenko, V. A. (1997). Phage Display. *Chemical Reviews, 97*(2), 391-410. doi: 10.1021/cr960065d

Staff. (2013). Trends in Health Care Cost Groth and The Role Of The Affordable Care Act: Office of the President of the United States.

Stafford, P., Cichacz, Z., Woodbury, N. W., & Johnston, S. A. (2014). Immunosignature system for diagnosis of cancer. *Proceedings of the National Academy of Sciences, 111*(30), E3072-E3080. doi: 10.1073/pnas.1409432111

Stafford, P., Halperin, R., Legutki, J. B., Magee, D. M., Galgiani, J., & Johnston, S. A. (2012). Physical Characterization of the 'Immunosignaturing Effect'. *Molecular & Cellular Proteomics*. doi: 10.1074/mcp.M111.011593

Stegmann, C. M., Lührmann, R., & Wahl, M. C. (2010). The Crystal Structure of PPIL1 Bound to Cyclosporine A Suggests a Binding Mode for a Linear Epitope of the SKIP Protein. *PLoS One, 5*(4), e10013. doi: 10.1371/journal.pone.0010013

Su, X.-z., Heatwole, V. M., Wertheimer, S. P., Guinet, F., Herrfeldt, J. A., Peterson, D. S., . . . Wellems, T. E. (1995). The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes. *Cell, 82*(1), 89-100. doi: http://dx.doi.org/10.1016/0092-8674(95)90055-1

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550. doi: 10.1073/pnas.0506580102

Sykes, K. F., Legutki, J. B., & Stafford, P. (2013). Immunosignaturing: a critical review. *Trends in biotechnology, 31*(1), 45-51.

Sykes, K. F., Legutki, J. B., & Stafford, P. (2013). Immunosignaturing: a critical review. *Trends Biotechnol, 31*(1), 45-51. doi: S0167-7799(12)00192-8 [pii]
10.1016/j.tibtech.2012.10.012 [doi]

Teng, G., & Papavasiliou, F. N. (2007). Immunoglobulin Somatic Hypermutation. *Annual Review of Genetics, 41*(1), 107-120. doi: doi:10.1146/annurev.genet.41.110306.130340

Vollmers, H. P., & Brändlein, S. (2007). Natural antibodies and cancer. *Journal of Autoimmunity, 29*(4), 295-302. doi: http://dx.doi.org/10.1016/j.jaut.2007.07.013

Wagner, B., Freer, H., Rollins, A., Garcia-Tapia, D., Erb, H. N., Earnhart, C., . . . Meeus, P. (2012). Antibodies to Borrelia burgdorferi OspA, OspC, OspF, and C6 Antigens as Markers for Early and Late Infection in Dogs. *Clinical and Vaccine Immunology, 19*(4), 527-535. doi: 10.1128/cvi.05653-11

Wagner, S., Hafner, C., Allwardt, D., Jasinska, J., Ferrone, S., Zielinski, C. C., . . . Breiteneder, H. (2005). Vaccination with a human high molecular weight melanoma-associated antigen mimotope induces a humoral response inhibiting melanoma cell growth in vitro. *The Journal of Immunology, 174*(2), 976-982.

Wang, L.-F., & Yu, M. (2004). Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Current drug targets, 5*(1), 1-15.

Yip, Y. L., & Ward, R. L. (1999). Epitope discovery using monoclonal antibodies and phage peptide libraries. *COMBINATORIAL CHEMISTRY AND HIGH THROUGHPUT SCREENING, 2*, 125-138.

Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2012). Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *New England Journal of Medicine, 367*(19), 1814-1820. doi: doi:10.1056/NEJMoa1211721

Zasloff, M. (2002). Antimicrobial peptides of multicellular organisms. *Nature, 415*(6870), 389-395.

APPENDIX A

COVARIATE MODULES FOR DIMENSIONALITY REDUCTION

Introduction

Batch to batch variation is a problem for microarray based assays. This has been true since RNA expression arrays were invented and has continued as the technology expands into protein and peptide based microarrays. Next generation sequencing has been touted as a solution to this problem, but this is only applicable to RNA or DNA profiling, as protein and peptide based assays still rely on microarrays. Further, there are over 1.2 million samples of microarray data in the Gene Expression Omnibus database (GEO) that remains underexploited due to this issue. Methods such as ComBat attempt to solve this problem, but these are unsuitable for use in blinded tests as they require class labels in addition to batch labels as inputs. A general approach is needed for doing meta-analyses of noisy array data, be it peptide, protein or DNA/RNA.

Idea

Arrays consist of many measurements, not all of which are independent. There are many hidden groups of features that vary together or do not vary much at all. The CIM330K array may have 330,000 features, but it does not have 330,000 independent features, so the measured information is likely much less than is immediately apparent. If we could group features into clusters or "modules" of features with high covariance, perhaps these groupings could be used to reduce batch variation.

Method

Finding groups of similar features reduces to a clustering problem. There are many ways to do this, all of which suffer from similar issues. The main question is how many clusters should be formed based on the data? Hierarchical clustering avoids this problem completely by constructing a tree structure to represent the relationship between two instances, however the tree must be cut at a subjective point in order to create discrete clusters. K-means makes this subjectivity explicit by requiring the user to select how many clusters the data should be split into.

Dirichlet mixture modeling, which we use in this experiment, attempts to solve this problem by assuming an infinite number of possible clusters, and assigning instances to a discrete subset of these. Even this method invites a certain degree of subjectivity in that it takes a single parameter α, known as the concentration parameter. If α is large then many clusters are formed, if it is small then fewer clusters are formed. A prior distribution can also be placed over α if there is a good reason to do so, but in this preliminary experiment α is simply set to 1. See Figure 1 for more details. Once clusters are found, the data must be reduced by computing the mean of the cluster measurements. In an experiment with n samples, m peptides and k identified modules (clusters) the resulting transformed matrix would be $n\ by\ k$. The value in the ith row and jth column corresponds to the mean measurement of peptides in cluster j for sample i.

We selected two batches under which the same samples were run. These were the CIM7-18 and CIM7-30. Each contained a group of nine common Dengue and Malaria samples, and also seven distracter samples unique to each batch for a total of 16 samples in each batch. The data from CIM7-18 was used to define the peptide modules.

In order to measure the effect of this transformation on batch effects, correlation ranks were used. Correlation coefficients for each of the nine samples occurring in both batches were computed against each of the 16 samples from the other batch. These coefficients were ranked from highest to lowest, and the rank of the true identical sample was recorded for each before and after transformation. In this way, improvement relative to other samples could be assessed, since the number of variables is greatly reduced after transformation which may affect absolute measures of correlation. These absolute measures were also recorded and compared. A paired T Test was used to assess whether improvement was statistically significant.

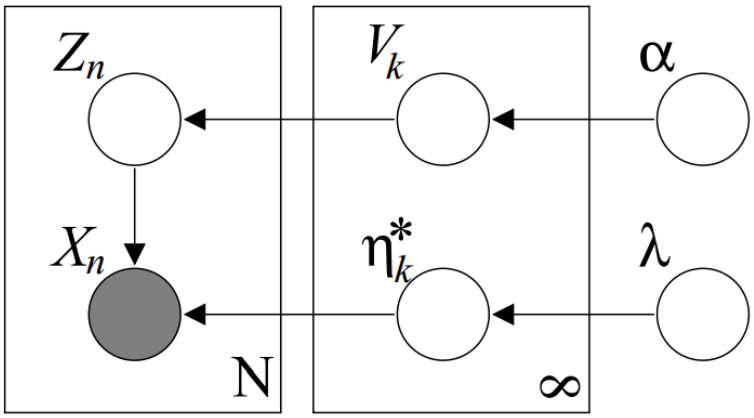


Figure A1.1: Plate diagram for general form of the Dirichlet process mixture model used in this experiment: This is a general form of the exponential family Dirichlet process mixture model. The parameters in realization of the model used in this experiment is as follows. The Dirichlet concentration hyperparameter $\boldsymbol{\alpha = 1}$, hyperparemeter $\boldsymbol{\lambda}$ refers to the way the Gaussian mixtures were initialized (diagonal covariance matrix, normally distributed means). $\mathbf{V_k}$ is an infinite dimensioned categorical distribution draw from the dirichlet process, $\boldsymbol{\eta_k^*}$ refers to a set of Gaussian parameters drawn according to $\boldsymbol{\lambda}$. $\mathbf{Z_n}$ is a cluster index 1 ... ∞ drawn from $\mathbf{V_k}$, which is used to select the parameters $\boldsymbol{\eta_k^*}$. Together these generate the data under the model, $\mathbf{X_n}$ which corresponds to a vector of measurements for a single peptide across all samples. These parameters can be fit with a Gibbs sampling process. See http://scikit-learn.org/stable/modules/generated/sklearn.mixture.DPGMM.html for more details on the actual process used.

Results and Discussion

Setting the number of mixture components to 100 yielded 59 modules (clusters) of peptides in the data tested. This produced a 59 x 16 matrix (modules x samples) for each of the two tested batches. Correlations between the 9 pairs of identical samples from each batch were compared before and after module-based transformation. There was a modest but significant improvement in correlation rank (P=0.016). The average rank improvement was 1.2. Correlations were greatly improved in the module transformed data, but this is not helpful information because the transformed data has far fewer dimensions (summarized in Figure A1.2). These results are very subtle, but the clustering method is also very crude. An optimization procedure could be designed seeking hyperparameters λ and α such that the rank improvement is maximized. A slight reformulation of the model used here could accomplish this.

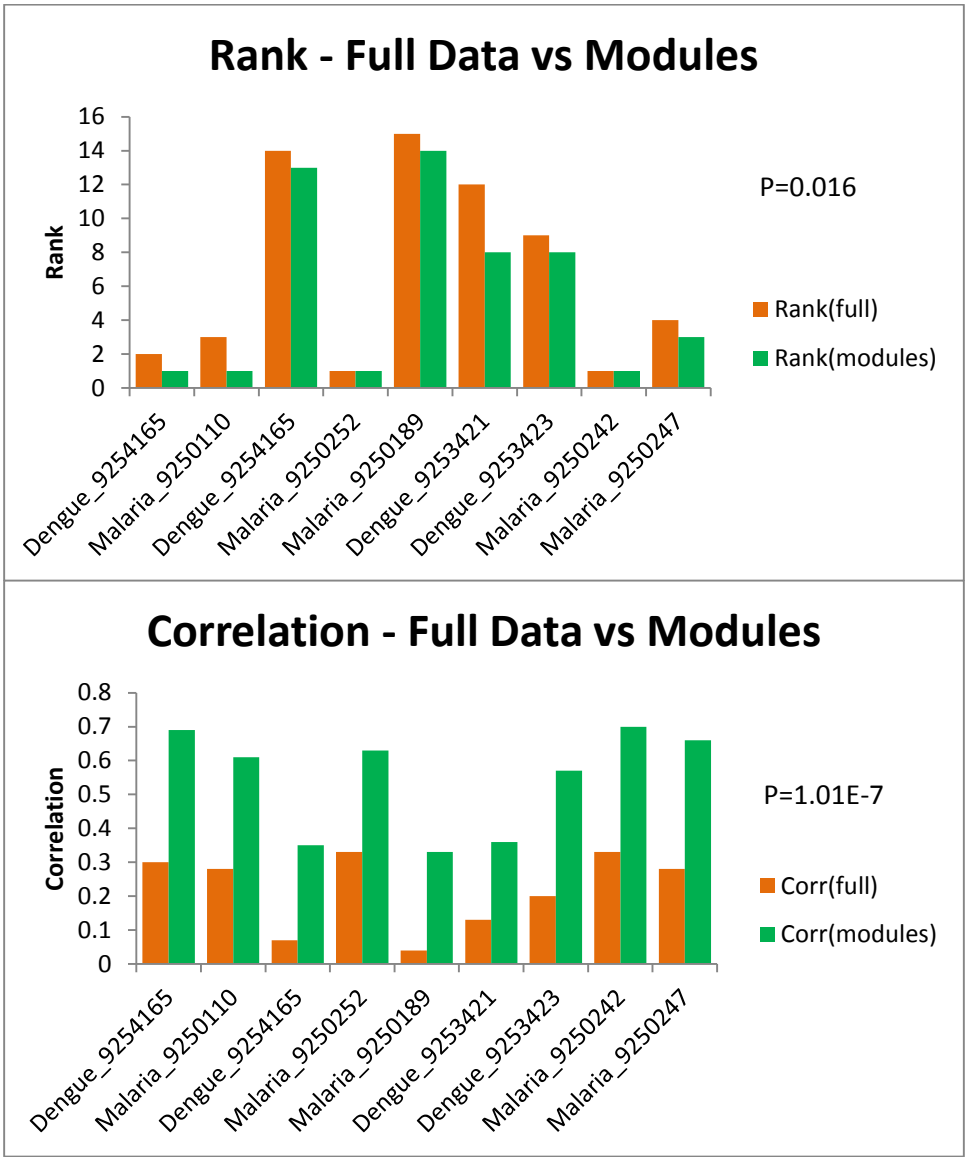


Figure A1.2: Correlation comparisons of module transformed versus raw data: Transformation showed a modest but significant improvement in rank, and a large improvement in correlation. Rank is likely more informative than correlation, as transformed correlation values are of much lower dimensionality than raw values and represent a vector of averages, thus are more likely to be higher by chance.

APPENDIX B

ALGORITHM FOR PARALLEL MASK DESIGN

Introduction

Recently there has been a lot of success developing mask based photolithographic synthesis of non-natural sequence peptide arrays (Legutki et al., 2014). In developing immunosignature based diagnostics it is important to cover as much sequence space as possible in order to provide a diverse surface on which antibodies can interact to create disease-specific patterns. This study examines the Artificial Immune System (a flavor of genetic algorithm) as an array generation method with the objective of maximizing measures of peptide diversity.

The process for achieving this optimal array is not straightforward for two reasons. First, there are multiple definitions of "complex surface" depending on the window length considered. For example, if 12-mers are considered the unit of diversity, then it is very easy to put down an array of 330,000 unique peptides since there are 16^{12} possible 12-mers in the current 16 amino acid synthesis scheme. This may be misleading, however, because these 12-mers could have hidden redundancies and biases at the subsequence level in the pentamer and hexamer windows of each peptide. What we are really interested in is maximizing the number of linear antibody binding sites on the surface, which are more poorly understood. Recent studies of linear epitopes using large combinatorial arrays have shed some light on this issue (Buus et al., 2012a), with the average monoclonal antibody requiring around 5 to 7 amino acids (but as few as 3) for strong binding **(Figure 1.7)**. This study attempts to maximize the number of unique 5-mers represented on the array.

The second reason an optimal array is difficult, even if an objective function can be found, is in the nature of the array synthesis. These are made using a mask based photolithographic process whereby a single amino acid is placed on a portion of array spots at each step. If one wants to make Q number of peptides of length M using N masks, one can imagine N masks stacked on top of each other with Q spots for potential holes on each. As one looks down through this stack of masks, there would be M holes in each spot **(Figure A2.1)** such that after all synthesis steps have been completed, the array consists entirely of peptides of length M. Due to this process, it is not straightforward to synthesize arbitrary peptides, since there are restrictions inherent in the combinatorial process of mask based synthesis. It is difficult to start from an optimal list of peptides and synthesize those exactly using masks.

In order to address this issue, we treated this as a combinatorial optimization problem and took inspiration from the natural system these arrays were designed to study: the clonal expansion and selection process that produces high affinity IgG antibodies.

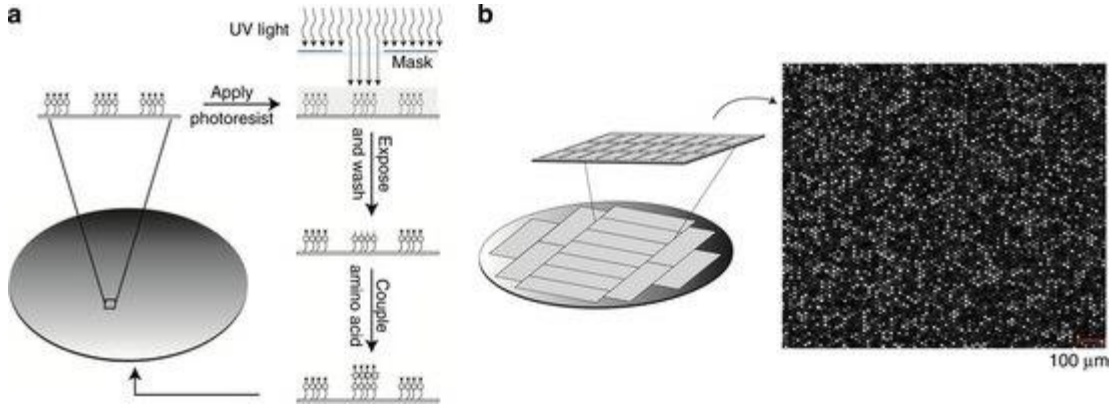


Figure A2.1: Overview of array synthesis (adapted with permission from (Legutki et al., 2014): The arrays considered here employ a mask based synthesis process, whereby amino acids are washed over the surface one at a time, and light flowing through holes in a mask directs the synthesis. By repeating this process with differing masks and amino acids, complex surfaces can be created.

Method

Genetic algorithms are intuitive and simple to implement, though their properties are difficult to analyze mathematically. They can be summarized in the following steps:

1. Initialize a population with random parameters.
2. Compute the objective function (the function we are trying to optimize) for each member of the population. The objective function in this case is simply the number of unique pentamers represented.
3. Take the "fittest" (highest objective function evaluation) individuals and use them for the next generation, and throw away the rest.
4. The fittest individuals produce a number of offspring to recapitulate the original population size. This can optionally involve a mating function for combining characteristics from multiple fit individuals, or one can simply "mutate" the parameters randomly to mimic VDJ recombination.
5. The new population shares characteristics of the parents (the fittest members of the previous generation) and additionally adds some random characteristics to explore more parameter space.
6. Repeat steps 2-6 until convergence.

The implementation used in this study more closely resembles bacterial replication or clonal selection, as it omits the mating function. The virtual array consists of 100,000 12-mer peptides. Thus each peptide has $12 - 5 + 1 = 8$ pentamer windows for a total of 800,000 possible unique binding sites. This is the upper bound on the number of unique pentamers that can be represented on this virtual array. The algorithm was tested against a naïve implementation that simply generated C arrays

using randomly assigned holes and amino acid orderings. C refers to the number of total "individuals" generated in the clonal selection process. These randomly generated arrays serve as a baseline control, indicating the best that can be done naively using the same number of computational steps. C in this test was 16 x 10000 = 160,000 (16 generations, 10,000 individuals per generation)

Unlike many optimization methods, the genetic algorithm (especially the clonal expansion variety) is highly amenable to parallelization. Each generation can be generated and the objective function evaluated on a separate core. Only the fitness selection step must be run in series, and this is an extremely rapid step.

Results

The algorithm was tested while varying the available number of masks between 40 and 240. For each mask length, the clonal expansion algorithm performed significantly better than naïve. At 40 masks, the naïve method generated 61,311 ± 11,879 unique 5-mers, while the clonal algorithm generated 120,589 sequences after 16 generations. The improvement was greatest at 120 masks, with the naïve method generating 502,900 ± 33,058 and the clonal method generating 697,743 pentamers, a 1.38 fold improvement. Parallelism was also successful, with speedup following a predicted near-linear trend. These results are summarized in **Figure A2.2**.

Discussion

Combinatorial optimization problems are difficult. The search space in this case is extremely large, and there are likely many local maxima. The objective function landscape is likely very "jagged" with a non-obvious and possibly unimportant global maximum. Our focus is on generating arrays that are good enough, not necessarily optimal, and for this task the clonal selection algorithm is clearly better than naïve methodology. That said, there are many other ways to accomplish this task including more rational top-down methods.

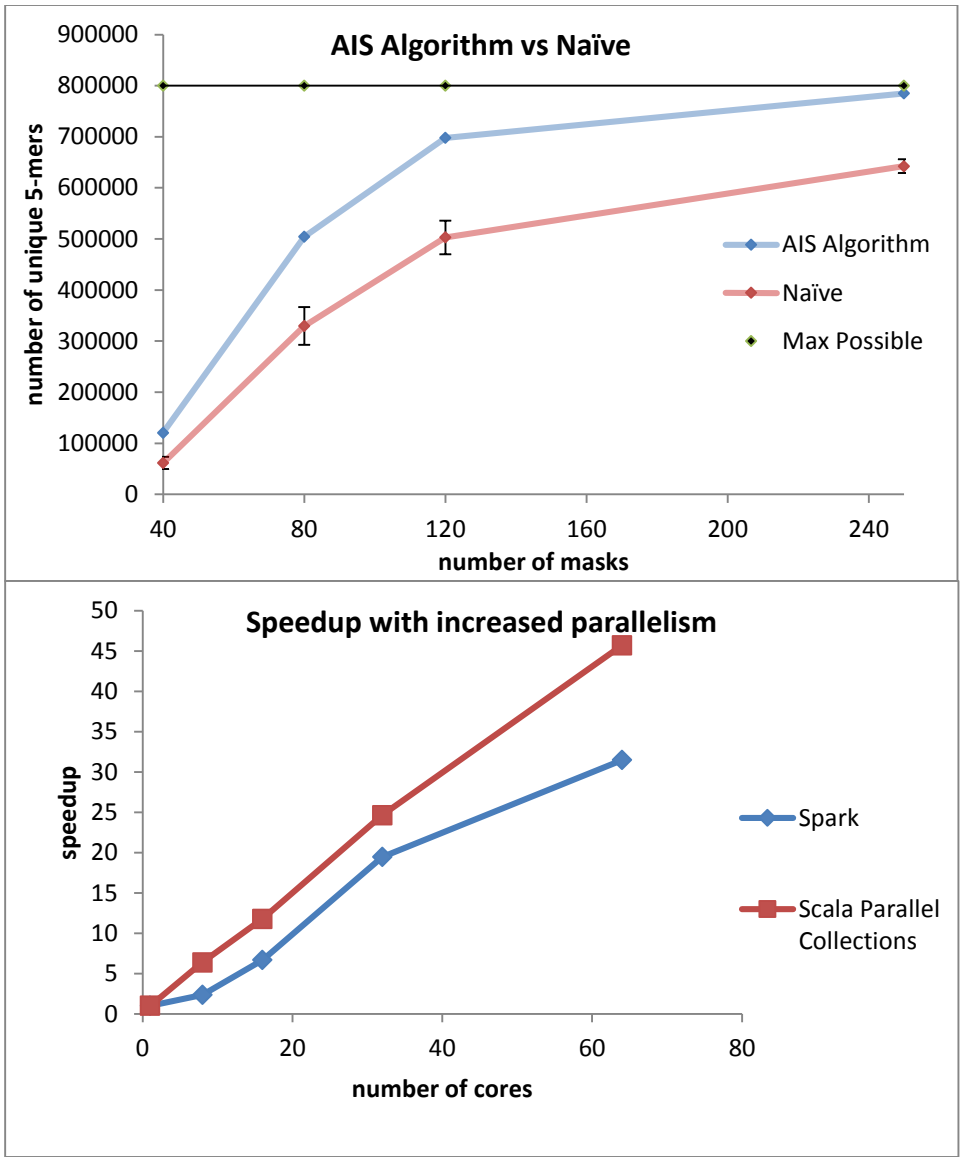


Figure A2.2: Clonal Expansion Algorithm (AIS) versus Naïve: Top shows clonal expansion algorithm (AIS) versus the naïve approach. At all mask lengths the AIS outperformed naïve methodology. It also scaled well, with speedup following a mostly linear trend (bottom). Spark refers to Apache Spark, a software package for running distributed operations on multiple cores or machines. Scala parallel collections refers to another, less scalable but more lightweight software library for parallelization. These tests were implemented using Scala on OpenJDK build 24.65.

APPENDIX C

MEASURING OFF RATES ON ARRAYS

Introduction

Understanding the kinetics of immunosignatures is of principal importance to the assay. One of the unknowns which was discussed in Chapter 4 is the effect of peptide density on the dissociation constant K_d. Peptide density is proportional to spotting concentration, which refers to the concentration of peptide included in the spotting plate used for array manufacturing. Lower concentrations deposit less peptide on the surface, though the exact final peptide surface density is unknown and very difficult to measure. In this experiment spotted two peptides at varying concentrations and assayed binding against approximately 25nM of p53Ab1 at flow rates of approximately 50uL/s in a 130uL flow chamber using to the method developed by Matt Greving (Matthew P Greving et al., 2010). One of the peptides contained the epitope for this antibody, RHSVV, and the other contained this sequence with one substitution RHSVK. Both peptides were 20 amino acids long and are listed below.

Peptide 1: RHSVVSGSG**RHSVV**SGSGSC

Peptide 2: EHHYPV**RHSVK**TQDKVMGSC

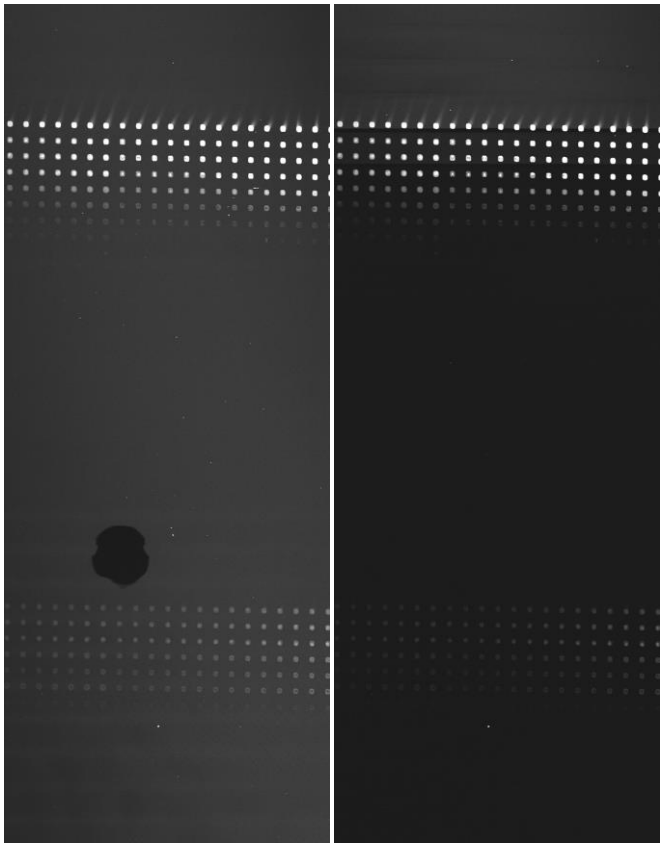
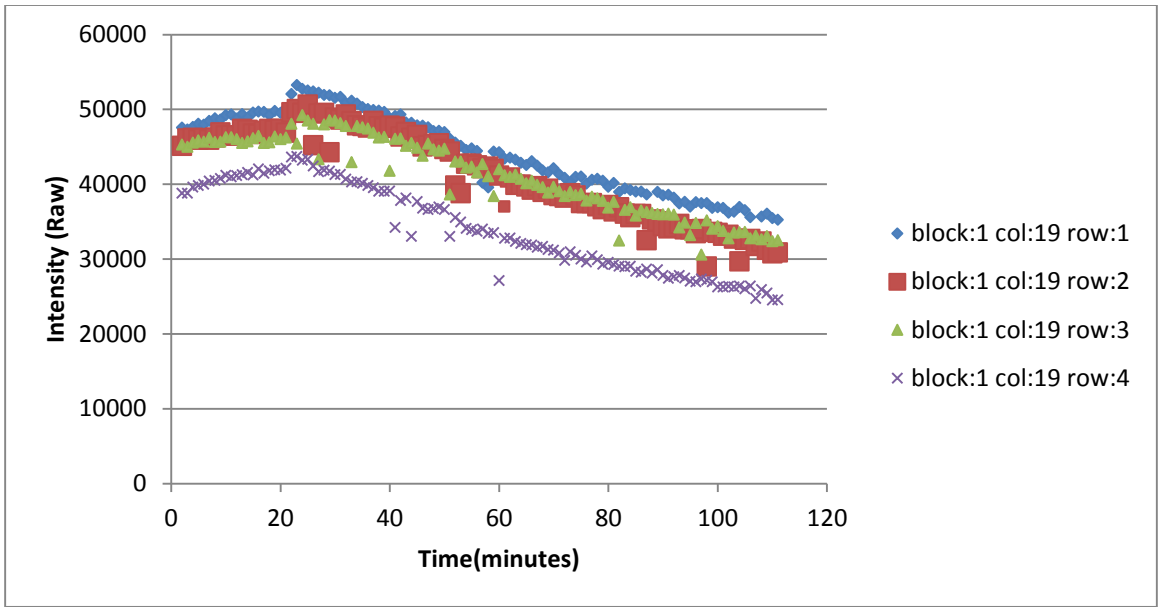
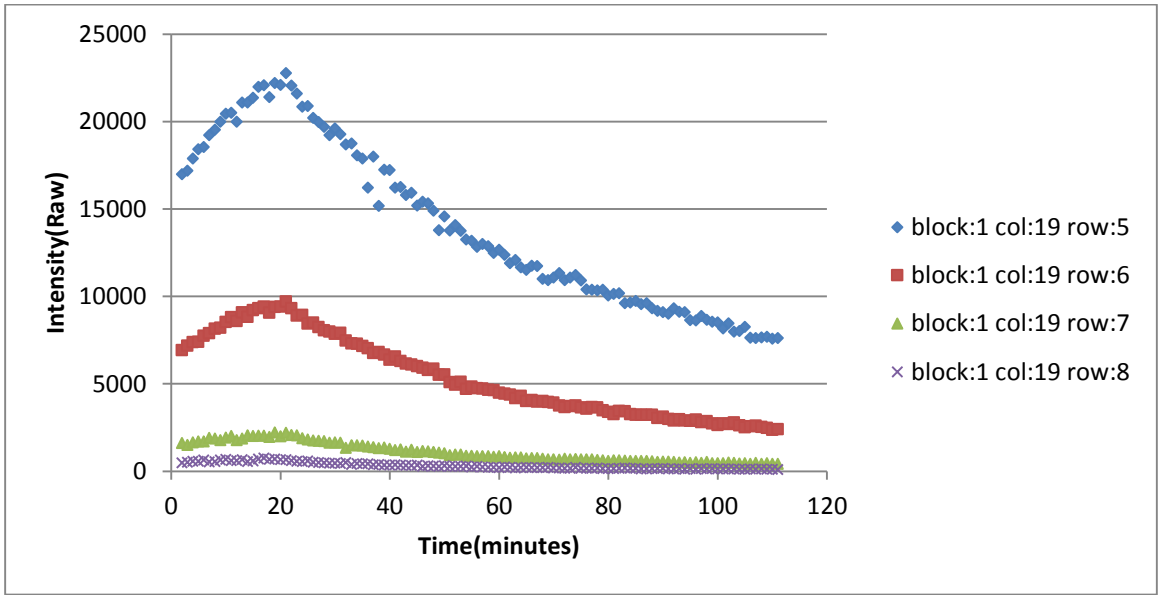


Figure AC.1: Flow cell Array Images
Using the method developed in (Matthew P Greving et al., 2010), I used the DNAScope to measure off rates of a single antibody with two different peptides. In the image above, the top set of peptides are repeated spots of RHSVVSGSG**RHSVV**SGSGSC and the bottom set is EHHYPV**RHSVK**TQDKVMGSC. Each row corresponds to a different concentration of spotted peptide. The top row of each block was spotted at 2mg/ml of peptide. Each subsequent row was serially diluted such that it had half the concentration of the row above. The antibody used was 25nM of p53Ab1, which was continually flowed over the surface of the array using a recirculating pump mechanism. The epitope for this antibody is RHSVV. The top spots contain this exactly and the bottom ones contain this with one substitution (RHSVK). The left figure shows a time slice while antibody is being flowed over the surface, and the right figure shows a different time slice during the wash step (buffer only). By taking several measurements over time, association and dissociation curves can be observed.



Concentrations of the rows 1 to 4: 2 mg/mL, 1 mg/mL, 0.5 mg/mL, 0.25 mg/mL



Concentrations of the rows 5 to 8: 0.125 mg/mL, 0.063 mg/mL, 0.031 mg/mL, 0.016 mg/mL

Figure AC.2: Association and Dissociation Curves of Selected Spots - Epitope: These are spots from the top block containing RHSVVSGSG**RHSVV**SGSGSC. At high peptide concentrations (top) off rate appears linear due to a saturation effect (likely optical). At low concentrations (bottom) a single component decay curve is observed.

Off rate estimation

A natural question from these data is how the off rate k_{off} is affected by decreased spotted peptide concentration. From **Figure AC.2** there are two concentrations: 0.125 mg/mL and 0.063 mg/mL where decent off rate exponential curves could be fit. Off rate units are in terms of fluorescence intensity, not standard molar or nanomolar quantities. That is to say, these are arbitrary units. Even so, it should be possible to tell from the exponsntial fits whether off rates are significantly different between these two spotting concentrations. Consider the following equation for exponential decay (a typical model for dissociation):

$$y = ce^{-kt}$$

Where y is the intensity at time *t*, *c* is a constant and *t* is time in minutes. This equation can be linearized as follows:

$$ln(y) = ln(c) - kx$$

This is a simple linear equation and can be fit with standard methods and used to estimate confidence intervals on k (the off rate in terms of arbitrary units). If there is a significant difference, then spotting density has an effect on off rate. Fitting the linearized equation to data from **Figure AC.2** yields the following estimates for k:

Spotting Conc.	*k estimate*	*Lower 95%*	*Upper 95%*	*R Squared*
Row 5: **0.125 mg/mL**	0.0116	0.0113	0.0120	0.983
Row 6: **0.063 mg/mL**	0.0148	0.0144	0.0153	0.982

Table AC.1: Off Rate Estimates at Two Spotting Concentrations: The confidence intervals for these two off rates (arbitrary units) do not overlap, indicating that spotting density could have an effect on off rates, though this effect is slight and these data are preliminary.

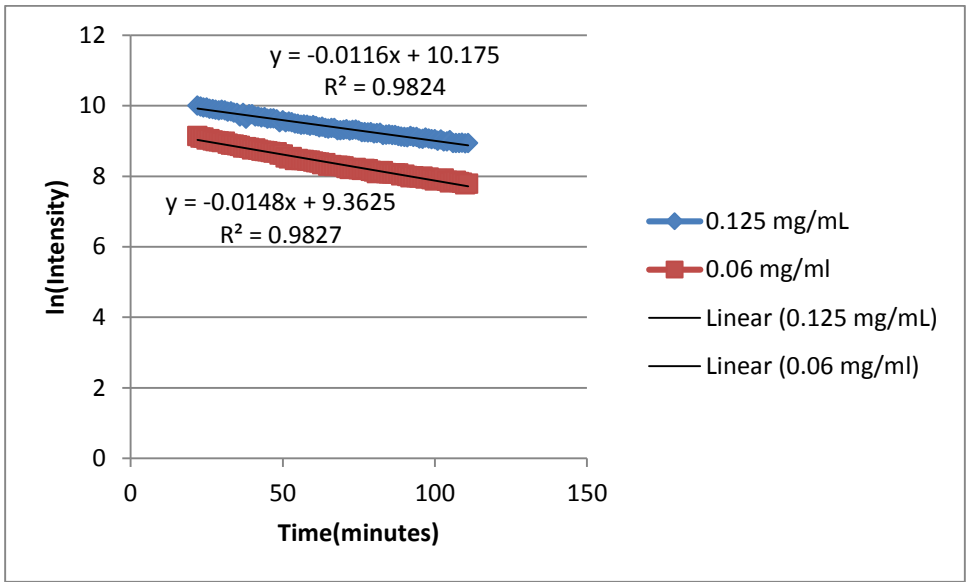
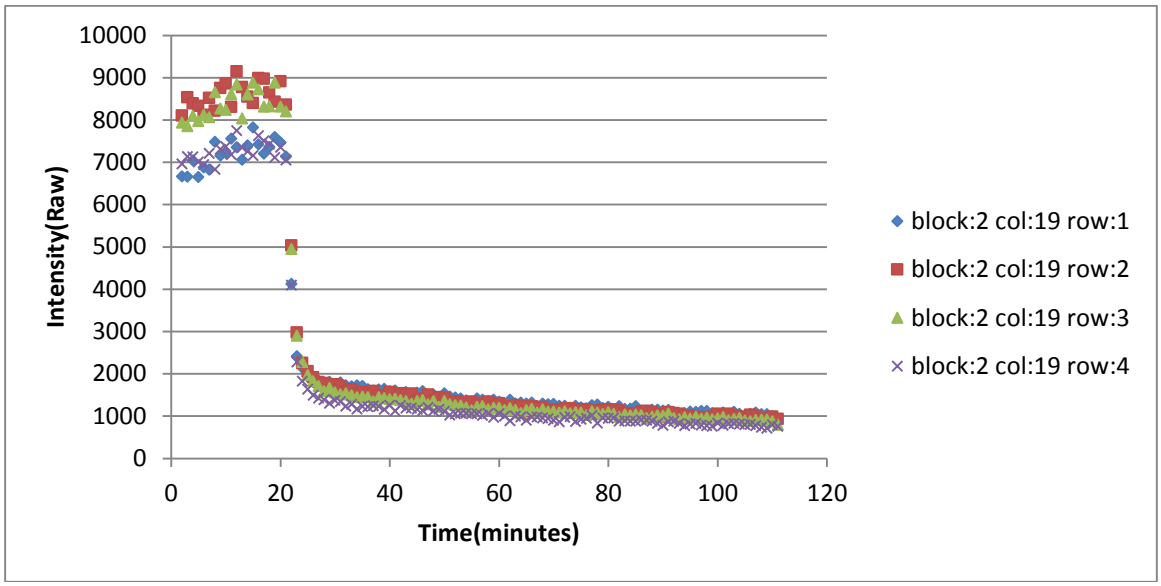
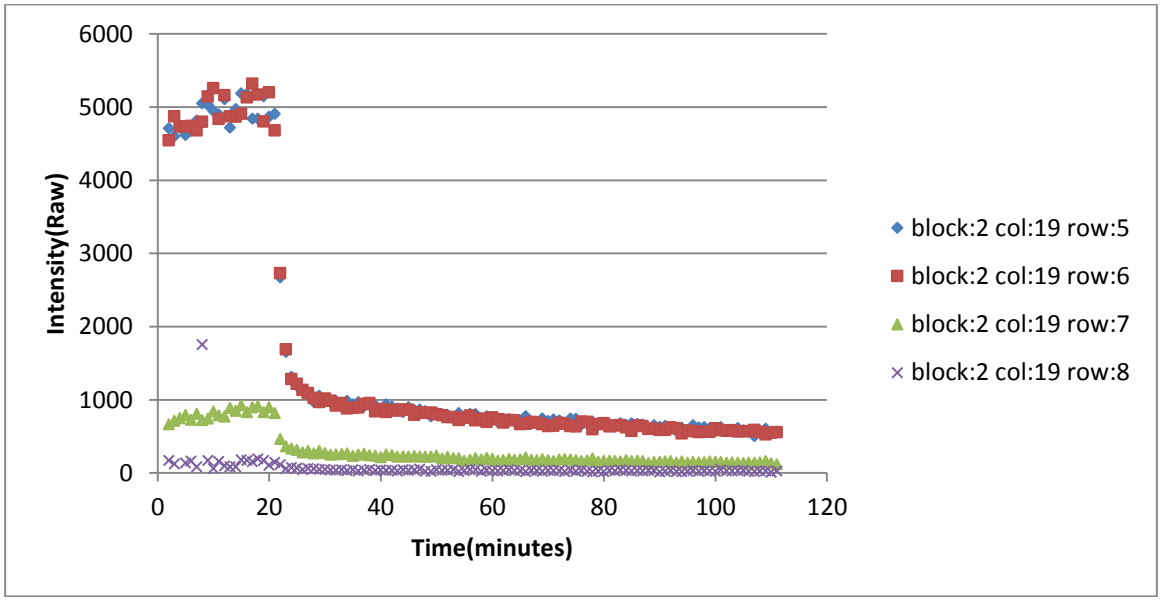


Figure AC.3: Linearized Off Rate Estimation: This is a graph of the linear fit shown in **Table AC.1**. The data follow a linear trend and a slight difference exists due to spotting density.



Concentrations of the rows (1 to 4): 2 mg/ml, 1 mg/ml, 0.5 mg/ml, 0.25 mg/ml



Concentrations of the rows (5 to 8): 0.125 mg/ml, 0.063 mg/ml, 0.031 mg/ml, 0.016 mg/ml

Figure AC.4: Association and Dissociation Curves of Selected Spots - Substitution:

These are spots from the bottom block containing EHHYPV**RHSVK**TQDKVMGSC with a single substitution from the epitope. At all peptide concentrations the off rate is rapid, but slows after an initial phase indicating a two component off rate. The

mechanism underlying this is unknown, but could indicate cooperative binding as discussed in experiments from (Halperin, 2011)

APPENDIX D

FIGURE PERMISSIONS

Figure 1.2 License

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Oct 26, 2014

This is a License Agreement between Joshua Richer ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number 3496520632912
License date Oct 26, 2014
Licensed content publisher Nature Publishing Group
Licensed content publication Nature Reviews Drug Discovery
Licensed content title The productivity crisis in pharmaceutical R&D
Licensed content author Fabio Pammolli, Laura Magazzini, Massimo Riccaboni
Licensed content date Jun 1, 2011
Volume number 10
Issue number 6
Type of Use reuse in a dissertation / thesis
Requestor type academic/educational
Format print and electronic
Portion figures/tables/illustrations
Number of figures/tables /illustrations
1
High-res required no
Figures Figure 1 | Trends in attrition rates of drug development projects.
Author of this NPG article no
Your reference number None
Title of your thesis / dissertation Non-Natural Sequence Peptide Microarrays for Diagnostics and Health Monitoring
Expected completion date Jan 2015
Estimated size (number of pages)
200
Total 0.00 USD
Terms and Conditions

Figure 1.3 Licence

10/26/2014 Rightslink Printable License
https://s100.copyright.com/AppDispatchServlet 1/3

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Oct 26, 2014

This is a License Agreement between Joshua Richer ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number 3496590748756
License date Oct 26, 2014
Licensed content publisher Nature Publishing Group
Licensed content publication
Nature
Licensed content title Broad and potent neutralization of HIV-1 by a gp41-specific human antibody
Licensed content author Jinghe Huang, Gilad Ofek, Leo Laub, Mark K. Louder, Nicole A. Doria-Rose, Nancy S. Longo, Hiromi Imamichi, Robert T. Bailer, Bimal Chakrabarti, Shailendra K. Sharma, S. Munir Alam, Tao Wang, Yongping Yang, Baoshan Zhang, Stephen A. Migueles, Richard Wyatt, Barton F. Haynes, Peter D. Kwong, John R. Mascola, Mark Connors
Licensed content date Sep 18, 2012
Volume number 491
Issue number 7424
Type of Use reuse in a dissertation / thesis
Requestor type academic/educational
Format print and electronic
Portion figures/tables/illustrations
Number of figures/tables/illustrations
1
High-res required no
Figures Figure 4 a and b
Author of this NPG article no
Your reference number None
Title of your thesis / dissertation
Non-Natural Sequence Peptide Microarrays for Diagnostics and Health Monitoring
Expected completion date Jan 2015
Estimated size (number of pages)
200
Total 0.00 USD

Figure 1.4 Licence

10/26/2014 Rightslink Printable License
https://s100.copyright.com/AppDispatchServlet 1/6

ELSEVIER LICENSE

TERMS AND CONDITIONS
Oct 26, 2014
This is a License Agreement between Joshua Richer ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.
All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.
Supplier Elsevier Limited
The Boulevard,Langford Lane
Kidlington,Oxford,OX5 1GB,UK
Registered Company
Number
1982084
Customer name Joshua Richer
Customer address 1925 S Shannon Dr
TEMPE, AZ 85281
License number 3496591323185
License date Oct 26, 2014
Licensed content publisher Elsevier
Licensed content
publication
Structure
Licensed content title Structure of CD40 Ligand in Complex with the Fab Fragment of a
Neutralizing Humanized Antibody
Licensed content author Michael Karpusas,Jodie Lucci,Janine Ferrant,Christopher Benjamin,Frederick R. Taylor,Kathy Strauch,Ellen Garber,Yen-Ming
Hsu
Licensed content date April 2001
Licensed content volume
number
9
Licensed content issue
number
4
Number of pages 9
Start Page 321
End Page 329
Type of Use reuse in a thesis/dissertation
Intended publisher of new
work
other
Portion figures/tables/illustrations
Number of
figures/tables/illustrations
1
Format both print and electronic

Are you the author of this
Elsevier article?
No

Will you be translating? No
Title of your thesis/dissertation
Non-Natural Sequence Peptide Microarrays for Diagnostics and Health Monitoring
Expected completion date Jan 2015
Estimated size (number of pages)
200
Elsevier VAT number GB 494 6272 12
Permissions price 0.00 USD
VAT/Local Sales Tax 0.00 USD / 0.00 GBP
Total 0.00 USD

APPENDIX E

PUBLICATIONS AND SUBMISSIONS

CHAPTER 2

This work was accepted for publication in Molecular and Cellular Proteomics on Nov. 3, 2014

Josh Richer, Stephen Albert Johnston, Phillip Stafford

Epitope Identification from Fixed-Complexity Random-Sequence Peptide Microarrays

All co-authors have granted permission for the inclusion of this work in this dissertation.

CHAPTER 3

This work is a manuscript under preparation.

Josh Richer, Xiao Wang, Phillip Stafford, Stephen Albert Johnston

Immunosignatures for Dengue Diagnostics

All co-authors have granted permission for the inclusion of this work in this dissertation.